TODO MUNDO MENTE

BIG DATA, NOVOS DADOS E O QUE A INTERNET NOS DIZ SOBRE QUEM REALMENTE SOMOS

ALTA BOOKS
EDITORA

SETH STEPHENS-DAVIDOWITZ

PREFÁCIO DE STEVEN PINKER

TODO MUNDO MENTE

TODO MUNDO MENTE

O que a internet e os
dados dizem sobre
quem realmente somos

SETH STEPHENS-DAVIDOWITZ

Todo Mundo Mente — O que a internet e os dados dizem sobre quem realmente somos
Copyright © 2018 da Starlin Alta Editora e Consultoria Eireli. ISBN: 978-85-508-0217-6

2018 – Edição revisada conforme o Acordo Ortográfico da Língua Portuguesa de 2009.

Publique seu livro com a Alta Books. Para mais informações envie um e-mail para autoria@altabooks.com.br

Obra disponível para venda corporativa e/ou personalizada. Para mais informações, fale com projetos@altabooks.com.br

Produção Editorial Editora Alta Books **Produtor Editorial** Thiê Alves	**Gerência Editorial** Anderson Vieira **Supervisão de Qualidade Editorial** Sergio de Souza	**Produtor Editorial (Design)** Aurélio Corrêa **Editor de Aquisição** José Rugeri j.rugeri@altabooks.com.br	**Marketing Editorial** Silas Amaro marketing@altabooks.com.br **Vendas Corporativas** Sandro Souza sandro@altabooks.com.br	**Vendas Atacado e Varejo** Daniele Fonseca Viviane Paiva comercial@altabooks.com.br **Ouvidoria** ouvidoria@altabooks.com.br
Equipe Editorial	Bianca Teodoro Christian Danniel	Ian Verçosa Illysabelle Trajano	Juliana de Oliveira Renan Castro	
Tradução Wendy Campos	**Copidesque** Carolina Gaio	**Revisão Gramatical** Thamiris Leiroza Priscila Gurgel	**Diagramação** Amanda Meirinho	**Capa** Aurélio Corrêa

Erratas e arquivos de apoio: No site da editora relatamos, com a devida correção, qualquer erro encontrado em nossos livros, bem como disponibilizamos arquivos de apoio se aplicáveis à obra em questão.

Acesse o site www.altabooks.com.br e procure pelo título do livro desejado para ter acesso às erratas, aos arquivos de apoio e/ou a outros conteúdos aplicáveis à obra.

Suporte Técnico: A obra é comercializada na forma em que está, sem direito a suporte técnico ou orientação pessoal/exclusiva ao leitor.

Dados Internacionais de Catalogação na Publicação (CIP)
Vagner Rodolfo CRB-8/9410

S832t Stephens-Davidowitz, Seth

Todo mundo mente: O que a internet e os dados dizem sobre quem realmente somos / Seth Stephens-Davidowitz ; traduzido por Wendy Campos. - Rio de Janeiro : Alta Books, 2018.

Tradução de: Everybody Lies
ISBN: 978-85-508-0217-6

Para mamãe e papai

SUMÁRIO

PREFÁCIO

Desde que os filósofos especularam sobre o “cerebroscópio”, um dispositivo mitológico capaz de mostrar os pensamentos de uma pessoa na tela, cientistas sociais buscam ferramentas para expor o funcionamento da natureza humana. Durante minha carreira como psicólogo experimental, diferentes técnicas entraram e saíram de moda, experimentei todas elas — escalas de classificação, tempos de reação, dilatação de pupila, neuroimagem funcional, até mesmo a utilização de implantes de eletrodos em pacientes com epilepsias, felizes em passar horas em um experimento de linguagem aguardando para ter uma convulsão.

No entanto, nenhum destes métodos oferece uma visão desobstruída da mente. O problema é uma barganha cruel. Pensamentos humanos são proposições complexas; ao contrário da leitura rápida de Woody Allen sobre *Guerra e Paz*, não pensamos dessa forma simplificada: “Era sobre uns russos.” Mas as proposições em todo seu intrincado e multidimensional esplendor são difíceis para um cientista analisar. Obviamente, quando as pessoas expõem suas emoções, compreendemos a riqueza de sua torrente de consciência; porém, os monólogos não são o conjunto de dados ideal para testar hipóteses. Por outro lado, se nos concentrarmos em mensurações facilmente quantificáveis, como o tempo de reação das pessoas a palavras ou a resposta cutânea a imagens, podemos calcular as estatísticas, mas reduziríamos a complexa textura da cognição a um único número. Mesmo as metodologias mais sofisticadas de neuroimagem só conseguem nos dizer como os pensamentos se disseminam no espaço 3-D, mas não no que consistem.

Como se a barganha entre a tratabilidade e a riqueza não fosse ruim o bastante, cientistas da natureza humana são atormentados pela Lei dos Pequenos Números — o nome atribuído por Amos Tversky e Daniel Kahneman para a falácia do pensamento que afirma que as características de uma população serão refletidas em qualquer amostra, não importa o quão pequena. Mesmo os cientistas mais “numéricos” têm, lamentavelmente, percepções deficientes sobre quantos sujeitos são realmente necessários em um estudo antes que se possa generalizar a partir de peculiaridades aleatórias e universalizar para todos os norte-americanos, e muito menos para todos os *Homo sapiens*. É sempre ainda mais incerto quando a amostra é colhida por conveniência, como, por exemplo, ao se oferecer um dinheiro extra para alunos de nossos cursos.

Este livro trata de uma maneira totalmente nova de estudar a mente. O Big Data de buscas de internet e outras respostas online não é um cerebroscópio, mas Seth Stephens-Davidowitz mostra que oferece uma visão sem precedentes da psique humana. Na privacidade de seus teclados, as pessoas confessam coisas das mais estranhas, às vezes — como em sites de encontros ou buscas por aconselhamento profissional —, porque geram consequências na vida real; em outras, exatamente porque não têm consequências: as pessoas são capazes de desabafar seus desejos ou medos sem uma pessoa real reagindo com espanto ou algo pior. De qualquer modo, as pessoas não estão apenas apertando um botão ou girando uma maçaneta, mas teclando trilhões de sequências de caracteres para descrever seus pensamentos em toda sua vastidão combinatória. Melhor ainda, deixam seus rastros digitais de uma forma fácil de reunir e analisar. São dados de pessoas de todas as esferas profissionais e sociais, que participam de experimentos não invasivos que variam os estímulos e organizam as respostas em tempo real e fornecem alegremente esses dados em números colossais.

Todo Mundo Mente é mais do que uma prova de conceito. De forma reiterada, minhas preconcepções sobre meu país e minha espécie foram subvertidas pelas descobertas de Stephens-Davidowitz. De onde vem o inesperado apoio a Donald Trump? Quando Ann Landers perguntou a seus leitores, em 1976, se eles se arrependiam de ter tido filhos, e se chocou ao descobrir que a resposta da maioria era sim, teria ela sido enganada por uma amostra autosselecionada e atípica? A internet é a culpada pela crise do final da década de 2010, chamada de "filtro bolha"? Quais os gatilhos dos crimes de ódio? As pessoas procuram piadas para ficar mais alegres? E embora goste de pensar que nada consegue me chocar, fiquei bastante espantado pelo que a internet revela sobre a sexualidade humana — incluindo a descoberta de que a cada mês um certo número de mulheres busca pela expressão "transando com bichos de pelúcia". Nenhum experimento que use tempo de reação, dilatação de pupila ou neuroimagem funcional jamais revelaria este fato.

Todos irão gostar de *Todo Mundo Mente*. Com curiosidade incansável e admirável sagacidade, Stephens-Davidowitz aponta um novo caminho para a ciência social do século XXI. Através dessa janela infinitamente fascinante sobre as obsessões humanas, quem precisa de um cerebroscópio?

— Steven Pinker, 2017

TODO MUNDO MENTE

INTRODUÇÃO

OS CONTORNOS DE UMA REVOLUÇÃO

Seguramente ele perderia, disseram.

Na prévia da eleição presidencial do Partido Republicano dos Estados Unidos em 2016, especialistas em eleições concluíram que Donald Trump não tinha chances. Afinal, Trump havia insultado diversos grupos minoritários. As pesquisas e seus analistas diziam que poucos norte-americanos aprovavam tais ofensas.

A maioria dos especialistas em pesquisas de opinião na época acreditava que Trump perderia a eleição geral. Muitos de seus prováveis eleitores disseram ter sido dissuadidos pelas visões e atitudes de Trump.

Mas na realidade havia algumas pistas de que Trump poderia de fato vencer tanto a prévia quanto a eleição — na internet.

Sou um especialista em dados de internet. Todos os dias, sigo os rastros digitais deixados pelas pessoas enquanto navegam na web. A partir dos botões ou teclas que clicam ou digitam, tento entender o que realmente querem, o que realmente fazem e quem realmente são. Vou lhe explicar como comecei esta trajetória incomum.

A história começa — e parece que isso foi há séculos — com a eleição presidencial de 2008 e uma questão muito debatida em ciência social: quão significativo é o preconceito racial nos Estados Unidos?

Barack Obama estava concorrendo como o primeiro candidato afro-americano de um grande partido. Ele venceu — com grande facilidade. E as pesquisas indicavam que a raça não era um fator determinante no modo de votar dos norte-americanos. O instituto Gallup, por exemplo, realizou inúmeras pesquisas antes e depois da primeira eleição de Obama. A conclusão? A grande maioria dos eleitores norte-americanos não se importava que Barack Obama fosse negro. Logo após as eleições, dois renomados professores da Universidade da Califórnia, em Berkeley, analisaram outras pesquisas baseadas em dados, usando técnicas de mineração de dados mais sofisticadas. Chegaram a conclusão similar.

E assim, durante o mandato de Obama, essas técnicas se transformaram em sabedoria convencional em muitas partes da mídia e searas do mundo acadêmico. As fontes utilizadas por cientistas sociais e de mídia por mais de 80 anos para entender o mundo nos diziam que a esmagadora maioria dos norte-americanos não se importava com o fato de Obama ser negro ao avaliar se deveria ou não ser presidente.

Os Estados Unidos, maculados por muito tempo pela escravidão e pelas leis de Jim Crow, pareciam finalmente ter parado de julgar as pessoas pela cor de sua pele. Isto parecia sugerir que o racismo dava seus últimos suspiros nos Estados Unidos. Na verdade, alguns estudiosos chegaram a declarar que vivíamos em uma sociedade pós-racial.

Em 2012, eu era um estudante de pós-graduação em economia perdido na vida, exaurido em meu campo de atuação, confiante, e até mesmo arrogante, de que tinha uma compreensão muito abrangente de como o mundo funcionava, de como as pessoas pensavam e com o que se importavam no século XXI. E quando se tratava de preconceito, eu me permitia acreditar, com base em tudo que li sobre psicologia e ciência política, que o racismo explícito era restrito a um pequeno porcentual de norte-americanos — a maioria composta por Republicanos conservadores, grande parte vivendo no extremo sul dos Estados Unidos.

Foi então que descobri o Google Trends.

O Google Trends, uma ferramenta lançada com pouco alarde em 2009, informa aos usuários com que frequência qualquer palavra ou frase foi pesquisada em locais diversos em diferentes momentos. Foi anunciado

como uma ferramenta de diversão — talvez permitindo que amigos discutam quais celebridades são mais populares ou a nova tendência da moda. As primeiras versões incluíam uma divertida frase de censura dizendo que as pessoas “não gostariam de escrever uma tese de doutorado” com os dados, o que imediatamente me motivou a escrever minha tese com eles.*

Na época, os dados de busca do Google não pareciam uma fonte adequada de informação para pesquisas acadêmicas “sérias”. Ao contrário das pesquisas, eles não foram criados como instrumento para ajudar a entender a psique humana. O Google foi criado para que as pessoas pudessem aprender sobre o mundo, não para que pesquisadores entendessem pessoas. Mas ocorre que os rastros que deixamos ao buscar conhecimento na internet são incrivelmente reveladores.

Em outras palavras, as buscas por informações são, por si só, informação. Quando e onde as pessoas buscam fatos, citações, piadas, pessoas, coisas ou ajuda, ao que se constata, pode nos dizer muito mais sobre o que realmente pensam, desejam, temem e fazem do qualquer um jamais imaginaria. Isto é especialmente verdade porque as pessoas não apenas fazem suas buscas, como também usam o Google para desabafar: “Odeio meu chefe”, “Estou bêbado”, “Meu pai me bateu”.

O ato cotidiano de digitar uma palavra ou frase em uma pequena caixa branca retangular deixa um pequeno rastro de verdade que, quando multiplicado por milhões, eventualmente revela profundas realidades. A primeira palavra que digitei no Google Trends foi “Deus”. Descobri que os estados norte-americanos que mais faziam buscas incluindo a palavra “Deus” eram Alabama, Mississippi e Arkansas — o chamado Cinturão da Bíblia. E essas pesquisas eram mais frequentes aos domingos. Nada disso era surpreendente, mas era intrigante que os dados de busca revelassem um padrão tão claro. Experimentei “Knicks” e descobri que é o termo mais pesquisado na cidade de Nova York. Outra obviedade. Então digitei meu nome e o Google Trends me informou: “Não há dados de pesquisa suficientes para exibir aqui.” Descobri, assim, que o Google Trends apenas fornece dados quando muitas pessoas fazem a mesma busca.

No entanto, o poder das buscas do Google não está em conseguir informar que Deus é popular no sul dos Estados Unidos, os Knicks, em Nova York ou que eu não sou popular em lugar algum. Qualquer pesquisa lhe informaria isso. O poder dos dados do Google está no fato de as pessoas “contarem” ao gigantesco mecanismo de busca o que não diriam a mais ninguém.

Por exemplo, sobre sexo (um assunto que investigo em detalhes mais adiante no livro). Pesquisas não são confiáveis para nos dizer a verdade sobre nossa vida sexual. Analisei os dados da General Social Survey, que é considerada uma das fontes de informação mais influentes e confiáveis sobre os comportamentos da sociedade norte-americana. De acordo com esta pesquisa, quando se trata de sexo heterossexual, mulheres dizem fazer sexo, em média, 55 vezes por ano, usando preservativo em 16% das vezes. Isto equivale a cerca de 1,1 bilhão de preservativos usados por ano. Mas homens heterossexuais dizem usar 1,6 bilhão de preservativos por ano. Estes números, por definição, teriam que ser iguais. Então, quem está dizendo a verdade, os homens ou as mulheres?

Ninguém, ao que se constata. De acordo com a Nielsen, uma empresa de mensuração e informação global que rastreia o comportamento de consumo, menos de 600 milhões de camisinhas são vendidas por ano. Então, todo mundo mente; a única diferença é o quanto.

A mentira é de fato amplamente disseminada. Homens que nunca foram casados alegam usar uma média de 29 preservativos por ano. Isto equivaleria a mais do que o total de preservativos vendidos nos Estados Unidos para pessoas casadas e solteiras. Pessoas casadas provavelmente também exageram na frequência com que fazem sexo. Em média, homens casados com menos de 65 anos dizem aos pesquisadores que fazem sexo uma vez por semana. Apenas 1% diz que passou o último ano sem sexo. Mulheres casadas relatam fazer menos sexo, mas não muito menos.

As buscas do Google nos dão uma visão muito menos animadora — e, eu defendo, muito menos precisa — do sexo no casamento. No Google, a principal reclamação sobre o casamento é a falta de sexo. Buscas por “casamento sem sexo” são 3,5 vezes mais comuns do que “casamento infeliz” e oito mais comuns do que “casamento sem amor”. Mesmo casais não casados reclamam com certa frequência de não fazer sexo. As buscas no Google por “relacionamento sem sexo” ficam em segundo lugar apenas para buscas por “relacionamentos abusivos”. (Estes dados, devo enfatizar, são todos apresentados anonimamente. O Google, é claro, não informa dados sobre buscas de qualquer indivíduo em particular.)

O retrato dos Estados Unidos criado pelas buscas no Google apresenta uma diferença impressionante da utopia pós-racial imaginada pelos pesquisadores. Lembro quando digitei pela primeira vez a palavra “nigger” [termo de cunho extremamente pejorativo usado para designar negros] no Google Trends. Podem me chamar de ingênuo. Mas considerando o quanto esta palavra é ofensiva, esperava sinceramente que a busca resultasse em um baixo volume de ocorrências. Cara, como estava errado. Nos Estados Unidos, a palavra “nigger” — ou seu plural, “niggers” — aparecia em praticamente o mesmo número de buscas que palavras como “enxaqueca”, “economista” e “Lakers”.

Imaginei que as buscas por letras de rap teriam influenciado os resultados. Não. A palavra usada em letras de rap é quase sempre “nigga(s)” [gíria derivada do termo nigger(s)]. Então, qual era a motivação dos norte-americanos que buscam pela palavra “nigger”? Frequentemente, eles procuravam por piadas para zombar afro-americanos. Na verdade, 20% das buscas com a palavra “nigger” também incluíam o termo “jokes” [piadas]. Outras buscas comuns continham expressões como “stupid niggers” [crioulos idiotas] e “I hate niggers” [odeio crioulos].

São milhões de buscas deste tipo todo ano. Um grande número de norte-americanos estava, na privacidade de seus lares, fazendo buscas terrivelmente racistas. Quanto mais pesquisava, mais perturbadora a informação se tornava.

Na noite da primeira eleição de Obama, quando a maioria dos comentários focava elogios a Obama e o reconhecimento da natureza histórica de sua eleição, praticamente uma em cada cem buscas no Google que incluíam a palavra “Obama” também continha “KKK” [Ku Klux Klan] ou “nigger(s)”. Pode não parecer muito, mas pense nos milhares de motivos de cunho não racista para pesquisar sobre esse jovem “forasteiro” com uma família encantadora que estava prestes a assumir o cargo mais poderoso do mundo. Na noite da eleição, buscas e assinaturas para o Stormfront, um site de nacionalismo branco surpreendentemente popular nos Estados Unidos, foram mais de dez vezes superiores ao normal. Em alguns estados, houve mais buscas por “nigger president” [presidente crioulo] do que por “first black president” [primeiro presidente negro].

Havia obscuridade e ódio escondido das fontes tradicionais, mas que se tornaram muito evidentes nas buscas que as pessoas fazem.

Essas buscam são difíceis de conciliar com uma sociedade em que o racismo seja um fator insignificante. Em 2012 conhecia Donald J. Trump principalmente como executivo e apresentador de *reality show*. Não tinha a menor ideia, assim como todo mundo, que quatro anos depois ele seria um sério candidato a presidente. Mas aquelas buscas terríveis não são difíceis de associar ao sucesso de um candidato que — em seus ataques aos imigrantes, em seus ódios e ressentimentos — frequentemente incitava as piores tendências nas pessoas.

As buscas no Google também nos indicam que muito do que pensávamos sobre a localização do racismo estava errado. As pesquisas e a sabedoria convencional imaginam o racismo moderno predominantemente localizado no sul dos Estados Unidos e majoritariamente entre Republicanos. Mas os lugares com as taxas de busca racistas mais altas incluem o norte do estado de Nova York, oeste da Pensilvânia, oeste de Ohio, área industrial de Michigan e área rural de Illinois, juntamente com Virgínia Ocidental, sul da Louisiana e Mississippi. A verdadeira divisão, os dados de busca no Google sugerem, não era do sul versus norte, mas sim do leste versus oeste dos Estados Unidos. Este tipo de coisa não é muito comum a oeste do Mississippi. E o racismo não se limitava aos Republicanos. Na verdade, buscas racistas não eram mais frequentes em lugares com alto porcentual de Republicanos do que em áreas dominantemente de Democratas. As buscas no Google, em outras palavras, ajudaram a desenhar um novo mapa do racismo nos Estados Unidos — e muito diferente do que você deve ter imaginado. Republicanos no sul são mais propensos a admitir o racismo. Mas muitos Democratas do norte têm atitudes semelhantes.

Quatro anos depois, este mapa se provaria significativo para explicar o êxito político de Trump.

Em 2012, eu usava o mapa do racismo que desenvolvi usando as buscas do Google para reavaliar exatamente o papel que a etnia de Obama desempenhou. Os dados eram claros. Em partes dos Estados Unidos com um alto número de buscas racistas, Obama obteve resultados substancialmente piores do que John Kerry, o candidato branco Democrata à presidência, quatro anos antes. A relação entre essas áreas não foi explicada por qualquer outro fator, incluindo níveis de escolaridade, idade, frequência a igrejas ou posse de arma. As buscas racistas não previram o baixo desempenho de qualquer outro candidato Democrata. Apenas o de Obama.

E os resultados sugeriam um grande impacto. Obama perdeu quase quatro pontos porcentuais no país inteiro apenas pelo racismo explícito. Isso foi muito superior ao que seria esperado com base em qualquer pesquisa. Barack Obama, é claro, foi eleito e reeleito presidente, com a ajuda de algumas condições bastante favoráveis para os Democratas, mas ele teve que superar muito mais do qualquer um, confiando nas fontes de dados tradicionais — e isso significava praticamente todo mundo —, havia percebido. Havia racistas suficientes para mudar o resultado da prévia ou da eleição geral em um ano que não fosse tão favorável para os Democratas.

Meu estudo foi inicialmente rejeitado por cinco jornais acadêmicos. Muitos dos pareceristas, me perdoem por expressar meu descontentamento, disseram que era impossível acreditar que tantos norte-americanos cultivassem um racismo tão ferrenho. Isso simplesmente não se encaixava com o que as pessoas estavam dizendo. Além disso, as buscas no Google pareciam um conjunto de dados um tanto bizarro.

Agora que testemunhamos a posse do Presidente Donald J. Trump, minhas descobertas parecem mais plausíveis.

Quanto mais estudava, mais descobria que o Google detém muitas informações que são ignoradas pelas pesquisas e que são úteis para compreender — dentre muitos e muitos assuntos — uma eleição.

Existem informações sobre quem realmente comparecerá às urnas. Mais da metade dos cidadãos que não votam diz nas pesquisas imediatamente antes de uma eleição que pretende votar, deturpando a estimativa de comparecimento às urnas, enquanto buscas no Google por "como votar" ou "onde votar" semanas antes de uma eleição podem prever com precisão quais as partes do país terão grande comparecimento às urnas.

Pode haver até mesmo informações sobre em quem irão votar. É possível realmente prever em qual candidato uma pessoa irá votar com base apenas em suas buscas? Claramente, não basta apenas pesquisar quais candidatos aparecem nas buscas com mais frequência. Muitas pessoas fazem buscas sobre um candidato porque gostam dele. Um número semelhante de pessoas faz buscas sobre um candidato porque o odeia. Sendo assim, Stuart Gabriel, professor de finanças na Universidade da Califórnia, Los Angeles, e eu descobrimos uma pista surpreendente sobre como as pessoas planejam votar. Um grande porcentual de buscas relacionadas às eleições contém uma questão envolvendo o nome de ambos os candidatos. Durante a eleição de 2016 entre Trump e Hillary Clinton, algumas pessoas buscaram por "pesquisa de votos Trump Clinton". Outras pesquisaram por pontos altos do "debate Clinton Trump". Na verdade, 12% das buscas com a palavra "Trump" também incluíam "Clinton". Mais de um quarto das buscas por "Clinton" também incluía "Trump".

Descobrimos que estas buscas aparentemente neutras nos dão algumas pistas sobre qual candidato a pessoa apoia.

Como? Pela ordem em que os nomes aparecem. Nossa pesquisa sugere que uma pessoa é significativamente mais propensa a digitar primeiro o nome do candidato que apoia em uma busca que inclua os nomes de mais de um.

Nas três eleições anteriores, o candidato que aparecia primeiro em mais buscas recebeu mais votos. E, mais interessante, a ordem em que foram pesquisados previu o posicionamento final de um determinado estado na eleição.

A ordem em que os candidatos são pesquisados também parece conter informações que as pesquisas ignoram. Na eleição de 2012 entre Obama e o Republicano Mitt Romney, Nate Silver, jornalista e gênio da estatística, previu o resultado em todos os cinquenta estados. Entretanto, descobrimos que em estados que listavam Romney antes de Obama nas buscas com mais frequência Romney realmente teve melhor desempenho do que Silver havia previsto. Em estados que listavam com mais frequência Obama antes de Romney, Obama se saiu melhor do que Silver previra.

Este indicador poderia conter informações que as pesquisas ignoram, pois os eleitores não estão mentindo para si mesmos nem desconfortáveis em revelar suas verdadeiras preferências para os pesquisadores. Embora declarassem estar indecisos em 2012, mas fazendo buscas por "pesquisas de votos Romney Obama", "debate Romney Obama" e "eleição Romney Obama" talvez estivessem de fato planejando votar em Romney.

Então, o Google previu Trump? Bem, ainda temos muito trabalho a fazer — e terei que contar com muito mais pesquisadores — antes que possamos saber como utilizar melhor seus dados para prever resultados de eleições. Esta é uma ciência nova e dispomos de estatísticas de poucas eleições. Certamente não estou afirmando que estamos no ponto — ou que chegaremos a ele — de poder realizar pesquisas públicas de opinião inteiramente como ferramenta para nos ajudar a prever eleições.

No entanto, definitivamente existiam presságios na internet, em muitos pontos, de que Trump poderia se sair melhor do que as pesquisas previam.

Durante a eleição geral, havia pistas de que o eleitorado pudesse ser favorável a Trump. Afro-americanos informaram nas pesquisas que compareceriam em massa às urnas para se opor a Trump. Mas as buscas no Google por informações sobre a votação em áreas predominantemente negras eram muito menores. No dia da eleição, Hillary foi prejudicada pelo baixo comparecimento dos negros.

Havia sinais até de que os eleitores indecisos escolheriam Trump. Gabriel e eu descobrimos que havia mais buscas por "Trump Clinton" do que por "Clinton Trump" em estados-chave no meio oeste, onde era esperado que Hillary vencesse. Na verdade, Trump deveu sua eleição ao fato de ter superado em muito o desempenho estimado pelas pesquisas nessas localidades.

Mas a principal pista, eu diria, de que Trump poderia se tornar um candidato vencedor — desde as prévias — era todo o racismo secreto que meu estudo sobre Obama havia revelado. As buscas no Google revelaram obscuridade e ódio em um número significativo de norte-americanos, que os estudiosos, por muitos anos, ignoraram. Os dados de busca revelaram que vivemos em uma sociedade muito diferente daquela em que acadêmicos e jornalistas, confiando em pesquisas, imaginavam que vivíamos. Eles revelaram um ódio perverso, assustador e amplamente disseminado que esperava por um candidato que lhe desse voz.

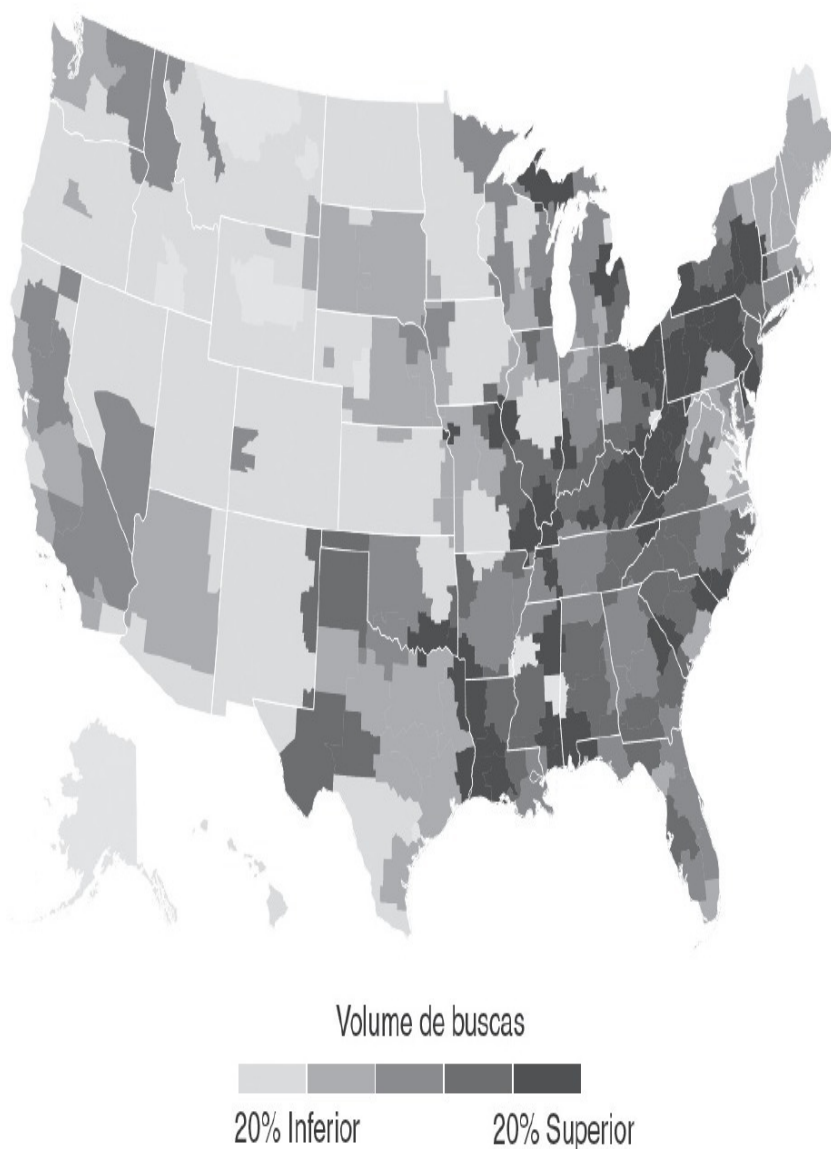
Com frequência, todo mundo mente — para si mesmo e para os outros. Em 2008, os norte-americanos informaram aos pesquisadores que não se importavam mais com raça. Oito anos depois, elegeram como presidente

Donald J. Trump, um homem que retuitou uma afirmação falsa de que os negros são responsáveis pela maioria dos homicídios de norte-americanos brancos, defendeu seus apoiadores por agredir um manifestante do movimento *Black Lives Matters* [Vidas de Negros Importam] em um de seus comícios e hesitou em repudiar o apoio de um ex-líder da Ku Klux Klan. O mesmo racismo oculto que prejudicou Barack Obama ajudou Donald Trump.

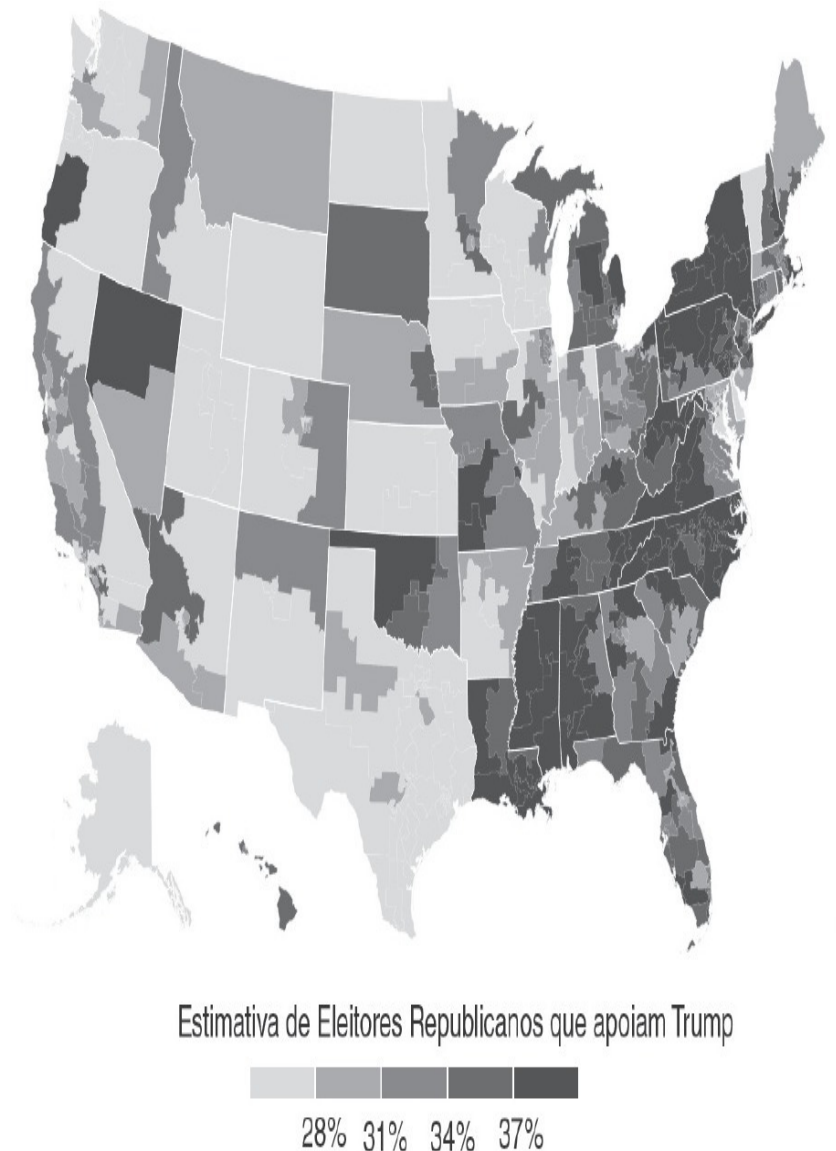
Nas pesquisas preliminares, Nate Silver declarou que as chances de Trump vencer eram virtualmente nulas. À medida que as prévias evoluíam e ficava cada vez mais claro que Trump detinha um apoio amplamente disseminado, Silver decidiu analisar os dados para tentar entender o que estava acontecendo. Como era possível que Trump se saísse tão bem?

Silver percebeu que as áreas em que Trump tinha melhor desempenho compunham um estranho mapa. Trump era mais popular em partes do nordeste e na área industrial do meio oeste, bem como no sul. Seus números eram notavelmente inferiores no oeste. Silver buscou variáveis para tentar explicar este quadro. Seria o desemprego? A religião? A posse de armas? As taxas de imigração? A oposição a Obama?

Silver descobriu que o único fator que se correlacionou adequadamente com o apoio a Donald Trump nas prévias do Partido Republicano foi minha descoberta quatro anos antes. As áreas que continham os maiores números de apoiadores de Trump eram as que faziam mais buscas no Google pelo termo “nigger”.

Taxa de Buscas Racistas

Apoio a Donald Trump na Prévia do Partido Republicano

Estimativa de Eleitores Republicanos que apoiam Trump

28% 31% 34% 37%

Passei quase todos os dias de meus últimos quatro anos analisando dados do Google. Esse período inclui uma posição como cientista de dados no Google, que me contratou depois de saber de minha pesquisa sobre racismo. E continuei a explorar os dados como colunista e jornalista de dados para o *New York Times*. As revelações continuavam surgindo. Doenças mentais; sexualidade humana; abuso infantil; aborto; propaganda; religião; saúde. Não são exatamente tópicos pequenos, e esse conjunto de dados, que não existia até algumas décadas, ofereceu perspectivas surpreendentes sobre todos eles. Economistas e outros cientistas sociais estão sempre à caça por novas fontes de dados, então serei direto: estou agora convencido de que as buscas no Google são o conjunto de dados mais importante jamais coletado sobre a psique humana.

Este conjunto de dados, porém, não é a única ferramenta que a internet oferece para a compreensão do mundo. Logo percebi que existem outras minas de ouro digitais. Baixei todo o conteúdo da Wikipédia, analisei perfis do Facebook e esquadrinhei o Stormfront. O PornHub, um dos maiores sites pornográficos da internet, me forneceu todos seus dados sobre buscas e visualização de vídeos de pessoas anônimas de todo o mundo. Em outras palavras, mergulhei profundamente no que hoje chamamos de Big Data. Além disso, entrevistei dezenas de pessoas — acadêmicos, jornalistas de dados e empreendedores — que também exploravam esses novos domínios. Muitos dos estudos dessas pessoas são discutidos neste livro.

Mas, primeiro, uma confissão: não darei uma definição precisa do que é Big Data. Por quê? Porque este é um conceito inerentemente vago. Quanto é "grande"? Small Data tem até 18.462 observações e Big Data, a partir de 18.463? Prefiro utilizar uma visão abrangente do que se qualifica como tal: apesar de a maioria dos dados que lido ser da internet, discuto outras fontes também. Vivemos em um período de explosão em quantidade e qualidade de todos os tipos de informação disponível. Grande parte das novas informações vêm do Google e das mídias sociais. Uma parcela disso é um produto da digitalização de informação que antes era guardada em arquivos e armários. Outra parte vem do aumento de recursos destinados a pesquisas de mercado. Alguns dos estudos discutidos neste livro não usam conjuntos de dados enormes, apenas empregam uma abordagem nova e criativa aos dados — essenciais em uma era de superabundância de informações.

Então por que exatamente o Big Data é tão poderoso? Pense em toda a informação que é disseminada pela internet em um determinado dia — temos um número, na verdade, para o quanto de informação existe. Em média, em um dia no início do século XXI, os seres humanos geraram 2,5 milhões de terabytes de dados.

E esses bytes são pistas.

Uma mulher entediada em uma tarde de quinta-feira faz uma busca no Google: "Piadas inocentes e engraçadas." Ela verifica o e-mail. Entra no Twitter. E busca no Google: "Piadas de crioulos."

Um homem está se sentindo triste. Ele pesquisa no Google "Sintomas de depressão" e "Histórias de depressão". Ele joga paciência.

Uma mulher vê o anúncio do noivado de uma amiga no Facebook. A mulher, que é solteira, bloqueia a amiga.

Um homem interrompe a busca no Google sobre a NFL e música rap para perguntar: "É normal sonhar que está beijando outro homem?"

Uma mulher clica em uma história do BuzzFeed mostrando "Os 15 gatos mais fofos".

Um homem vê a mesma história sobre gatos. Mas em sua tela aparece "Os 15 gatos mais adoráveis". Ele não clica.

Uma mulher busca: "Meu filho é um gênio?"

Um homem pesquisa no Google: "Como fazer minha filha perder peso?"

Uma mulher está de férias com as seis melhores amigas. Todas as amigas não param de dizer o quanto estão se divertindo. Ela entra no Google escondido e busca por: "Solidão longe do marido."

Um homem, o marido da mulher no exemplo anterior, está de férias com seus seis melhores amigos. Ele digita no Google: "Sinais de que sua mulher o está traindo."

Parte desses dados incluirá informações que nunca seriam admitidas por qualquer outro meio para outra pessoa. Se juntarmos todos eles, mantivermos o anonimato para garantir que nunca ficaremos sabendo sobre os medos, desejos e comportamentos de qualquer indivíduo específico, e adicionarmos um pouco de ciência de dados, começaremos a extrair uma nova perspectiva sobre os seres humanos — seus comportamentos, desejos, essências. Na verdade, correndo o risco de parecer pretensioso, chego a acreditar que os novos dados cada vez mais disponíveis em nossa era digital expandirão radicalmente nossa compreensão da humanidade. O microscópio nos mostrou que há mais em uma gota de água de um lago do que vemos. O telescópio nos mostrou que existe mais no céu noturno do que pensamos que vemos. E os novos dados digitais agora nos mostram que existe mais na sociedade humana do que pretendemos entender. Eles podem ser o microscópio ou o telescópio de nossa era — possibilitando percepções importantes e até revolucionárias.

Existe outro risco em fazer tais declarações — não apenas de parecer arrogância, mas também modismo. Muitas pessoas fazem grandes alegações sobre o poder do Big Data. Mas não têm provas relevantes.

Isso inspirou os céticos do Big Data, que também são muitos, a repudiar a busca por conjuntos de dados maiores. "Não estou dizendo que não há informação em Big Data", escreveu Nassim Taleb, estatístico e escritor de artigos. "Há muita informação. O problema — a questão central — é que a agulha vem em palheiros cada vez maiores."

Um dos principais objetivos deste livro, então, é fornecer as provas que faltam do que pode ser feito com Big Data — como podemos achar as agulhas, se preferir, em palheiros cada vez maiores. Espero fornecer exemplos suficientes de Big Data oferecendo novas perspectivas da psicologia e do comportamento humano para que você comece a ver os contornos de algo verdadeiramente revolucionário.

"Espera aí, Seth", você pode dizer agora mesmo. "Você está me prometendo uma revolução. Está se derramando em elogios sobre esses novos e grandes conjuntos de dados. Mas até agora, usou todos esses incríveis, maravilhosos, impressionantes e inovadores dados para dizer basicamente duas coisas: existe muito racismo nos Estados Unidos e as pessoas, especialmente os homens, exageram na frequência com que fazem sexo."

Admito; às vezes, os novos dados apenas confirmam o óbvio. Se você acha que essas descobertas são patentes, espere até chegar ao Capítulo 4, onde mostro provas claras e incontestáveis a partir de buscas no Google de que

homens têm enorme preocupação e inseguranças — ora, ora, que surpresa — com o tamanho de seus pênis.

Existe, eu diria, algum valor em provar coisas de que já se suspeitava, mas que tinha poucas evidências. Suspeitar de algo é uma coisa. Provar é outra. Mas se tudo o que o Big Data pudesse fazer fosse confirmar suas suspeitas, não seria revolucionário. Felizmente, o Big Data pode fazer muito mais do que isso. De forma reiterada, os dados me mostram que o mundo funciona de maneira exatamente oposta à que eu imaginava. Vejam alguns exemplos ainda mais surpreendentes.

Você pode pensar que as maiores causas do racismo são a insegurança e a vulnerabilidade econômica. Então deve naturalmente suspeitar que quando as pessoas perdem seus empregos, o racismo aumenta. Mas, na verdade, nem as buscas racistas nem as assinaturas no Stormfront aumentam quando o desemprego se eleva.

Pode pensar que a ansiedade é maior em grandes cidades, com níveis de escolaridade mais altos. O estereótipo urbano e neurótico é famoso. Mas as buscas no Google que refletem ansiedade — tais como "sintomas de ansiedade" ou "ajuda para ansiedade" — tendem a ser superiores em lugares com níveis de educação mais baixos, rendas médias menores e em que a maior parte da população vive em áreas rurais. Existem taxas de busca mais altas relacionadas à ansiedade em áreas rurais, no norte do estado de Nova York, do que na cidade de Nova York.

Pode imaginar que um ataque terrorista que mata dezenas ou centenas de pessoas é automaticamente acompanhado por ansiedade maciça e disseminada. O terrorismo, por definição, implica em infundir uma sensação de terror. Analisei as buscas no Google que refletissem ansiedade. Testei o quanto aumentaram em um país nos dias, semanas e meses seguintes a cada grande ataque terrorista na Europa e nos Estados Unidos desde 2004. Então, em média, em quanto as buscas relacionadas à ansiedade aumentaram? Elas não aumentaram. Nada.

Você pode imaginar que as pessoas buscam por piadas com mais frequência quando estão tristes. Muitos dos grandes pensadores da história alegaram que recorremos ao humor como forma de aliviar a dor. O humor há tempos é imaginado como um jeito de lidar com frustrações, dor e inevitáveis decepções da vida. Como Charlie Chaplin disse: "O riso é um tônico, um alívio, uma pausa que permite atenuar a dor."

Entretanto, as buscas por piadas são menores às segundas-feiras, o dia que as pessoas relatam que estão mais infelizes. São menores em dias nublados e chuvosos. E despencam depois de uma grande tragédia, como quando duas bombas mataram três pessoas e feriram centenas durante a Maratona de Boston, em 2013. As pessoas, na verdade, são mais propensas a procurar piadas quando a vida está bem do que quando não está.

Às vezes, um novo conjunto de dados revela comportamentos, desejos ou preocupações que eu nunca teria considerado. Existem inúmeras inclinações sexuais que se enquadram nesta categoria. Por exemplo, você sabia que na Índia a busca mais comum começada com "meu marido quer…" é "meu marido quer que eu o amamente"? Este comentário é muito mais comum na Índia que em outros países. Além do mais, buscas pornográficas por imagens de mulheres amamentando homens são quatro vezes maiores na Índia e em Bangladesh do que em qualquer outro país do mundo. Certamente eu nunca teria suspeitado disso antes de ver os dados.

Além do mais, enquanto o fato de os homens serem obcecados pelo tamanho de seus pênis pode não ser tão surpreendente, a maior insegurança das mulheres em relação ao próprio corpo, segundo as buscas no Google, é bastante inusitada. Com base nesses novos dados, a preocupação feminina equivalente ao tamanho do pênis nos homens parece ser — pausa dramática para criar suspense — com odores em suas vaginas. Mulheres fazem quase o mesmo número de buscas expressando preocupação sobre os próprios genitais que os homens em relação aos deles. E a maior preocupação expressada por mulheres é o odor — e como podem melhorá-lo. Certamente eu não sabia disso antes de ver os dados.

Às vezes os novos dados revelam diferenças culturais que nunca vislumbrei. Um exemplo: as muitas formas diferentes que homens em todo o mundo reagem ao fato de as esposas estarem grávidas. No México, as maiores buscas sobre "minha esposa grávida" incluem "frases de amor para minha esposa grávida" e "poemas para minha esposa grávida". Nos Estados Unidos, as principais buscas abrangem "minha esposa está grávida e agora" e "minha esposa está grávida, o que eu faço".

Mas este livro é mais do que uma coleção de fatos estranhos ou estudos isolados, embora exponha muitos casos assim. Como essas metodologias são muito novas e irão se tornar cada vez mais poderosas, mostrarei algumas ideias sobre como funcionam e o que as torna tão revolucionárias. Também reconhecerei as limitações do Big Data.

Um pouco da empolgação pelo potencial revolucionário dos dados é inapropriada. A maioria dos apaixonados pelo Big Data se entusiasma com a imensidão que eles podem atingir. Esta obsessão pelo tamanho dos conjuntos de dados não é nova. Antes do Google, Amazon e Facebook, antes que a expressão "Big Data" existisse, houve uma convenção realizada em Dallas, Texas, sobre "Conjuntos de Dados Grandes e Complexos". Jerry Friedman, professor de estatística em Stanford, que foi meu colega quando trabalhei na Google, se recorda da convenção de 1977. Um afamado estatístico fez uma palestra. Ele explicara que havia acumulado a quantidade impressionante e surpreendente de cinco gigabytes de dados. O segundo eminente estatístico a palestrar dissera: "O último palestrante tinha gigabytes. Isso não é nada. Eu tenho terabytes." A ênfase da palestra, em outras palavras, era sobre quanta

informação é possível acumular, não o que se esperava fazer com ela ou a quais questões planejava responder. “Achei isso divertido, na ocasião”, disse Friedman, “que a coisa que supostamente deveria nos impressionar era o tamanho do conjunto de dados. Isso ainda acontece.”

Muitos cientistas de dados hoje acumulam gigantescos conjuntos de dados e nos mostram coisas de pouca importância — ou seja, que os Knicks são populares em Nova York. Inúmeros negócios se afogam em dados. Eles possuem muitos terabytes, mas poucas percepções importantes. O tamanho de um conjunto de dados, acredito, é frequentemente superestimado. Existe uma explicação sutil, mas importante, para isso. Quanto maior a consequência, menor o número de observações necessárias para a ver. Só é preciso tocar o fogão quente uma vez para perceber que é perigoso. Pode ser necessário beber café milhares de vezes para determinar se ele lhe causa ou não enxaqueca. Qual lição é mais importante? Claramente, o fogão quente, que, em razão da intensidade de seu impacto, é aprendida tão rápido, com tão poucos dados.

Na verdade, as companhias mais inteligentes de Big Data frequentemente reduzem seus dados. No Google, as maiores decisões são baseadas em apenas uma minúscula amostra de todos seus dados. Você nem sempre precisa de uma tonelada de dados para descobrir percepções importantes. Precisa dos dados certos. Uma das principais razões pelas quais as buscas no Google são tão valiosas não é o fato de haver muitas; é o fato de as pessoas serem sinceras nelas. Todo mundo mente para amigos, parceiros, médicos, pesquisas e para si mesmo. Mas no Google as pessoas compartilham informações embaraçosas, sobre, entre outras coisas, casamentos sem sexo, problemas de saúde mental, inseguranças e animosidade contra pessoas negras.

Mais importante, para extrair percepções de Big Data, é preciso fazer as perguntas corretas. Assim como não podemos apontar um telescópio aleatoriamente para o céu noturno e encontrar Plutão, não se pode baixar uma enorme quantidade de dados e esperar que revelem segredos da natureza humana para você. É preciso olhar em lugares promissores — as buscas no Google que começam por “meu marido quer…” na Índia, por exemplo.

Este livro demonstrará qual é a melhor forma de usar Big Data e explicar em detalhes por que dados podem ser tão poderosos. E ao longo do caminho, você também aprenderá o que eu e outros analistas já descobrimos com eles, incluindo:

› Quantos homens são gays?
› Propaganda funciona?
› Por que o *American Pharoah* é um excelente cavalo de corrida?
› A mídia é tendenciosa?
› Atos falhos ou lapsos freudianos são reais?
› Quem sonega impostos?
› A universidade que você frequenta importa?
› Você é capaz de vencer o mercado de ações?
› Qual é o melhor lugar para criar os filhos?
› O que faz uma história viralizar?
› Sobre o que devemos falar no primeiro encontro se quisermos um segundo?

… e muito, muito mais.

Mas, antes, precisamos discutir uma questão mais elementar: por que precisamos dos dados? E, para isso, vou lhe apresentar minha avó.

* O Google Trends tem sido a fonte de grande parte de meus dados. Entretanto, como a ferramenta só permite a comparação da frequência relativa de diferentes buscas, mas não relata o número absoluto de qualquer busca específica, normalmente complemento os dados com o Google AdWords, que informa com que frequência cada busca é feita. Na maioria dos casos, fui capaz de aprimorar o resultado com a ajuda de meu próprio algoritmo baseado no Trends, que descrevo em minha tese, “Essays Using Google Data” [“Artigos Usando Dados do Google”, em tradução livre] e em meu artigo no *Journal of Public Economics* [Jornal de Economia Pública, em tradução livre], “The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data” [“O Custo do Ódio Racial para um Candidato Negro: Provas Usando os Dados da Busca do Google”, em tradução livre]. A tese, um link para o artigo e uma explicação completa dos dados e códigos usados em toda a pesquisa original apresentada neste livro estão disponíveis em meu site: sethd.com [conteúdo em inglês].

PARTE I

DATA, BIG E SMALL

1

SUA INTUIÇÃO FALHA

Se você tem 33 anos e frequenta alguns eventos de família seguidos sem uma namorada ou namorado, com certeza terá que enfrentar o assunto da escolha de um parceiro. E praticamente todo mundo terá uma opinião a oferecer.

"Seth precisa de uma garota maluca, como ele", diz minha irmã.

"Você está doida! Ele precisa de uma garota normal, para o equilibrar", fala meu irmão.

"Seth não é maluco", defende minha mãe.

"Você está louca! É claro que Seth é maluco", rebate meu pai.

De repente, minha avó, tímida e de fala suave, que estava quieta durante todo o jantar, resolve se pronunciar. As vozes nova-iorquinas agressivas silenciaram-se, e todos os olhos se voltaram para a pequena senhora de cabelos curtos e louros e uma voz que ainda revela traços de um sotaque do leste europeu. "Seth, você precisa de uma boa garota. Não muito bonita. Muito inteligente. Gentil com as pessoas. Sociável, para que você saia mais. Com senso de humor, pois você tem um ótimo senso de humor."

Por que o conselho dessa senhora mereceu tamanha atenção e respeito de minha família? Bem, minha avó, aos 87 anos, já presenciou mais eventos que qualquer outra pessoa na mesa. Ela observou muito mais casamentos, muitos que deram certo e muitos que não deram. E ao longo de décadas, ela catalogou as qualidades que possibilitam o sucesso de um relacionamento. Na mesa do jantar, para este assunto, minha avó tinha o maior número de pontos de dados. Minha avó é Big Data.

Neste livro, quero desmistificar a ciência de dados. Gostem ou não, os dados desempenham um papel cada vez mais importante em nossas vidas — e esse papel será ainda maior. Jornais agora têm seções inteiras dedicadas a dados. Empresas têm equipes com a tarefa exclusiva de analisar seus dados. Investidores oferecem dezenas de milhões de dólares a startups se forem capazes de armazenar mais dados. Mesmo que você nunca aprenda a executar uma regressão ou calcular um intervalo de confiança, vai se deparar com muitos dados — nas páginas que lê, em reuniões de negócios, nas fofocas que ouve em volta dos bebedouros.

Muitas pessoas estão angustiadas com esse crescimento. São intimidadas pelos dados, ficam perdidas e confusas no mundo dos números. Pensam que um entendimento quantitativo do mundo é só para alguns poucos prodígios especialistas em usar o lado esquerdo do cérebro, não para elas. Assim que se deparam com números, imediatamente viram a página, terminam a reunião, mudam de assunto.

No entanto, estou há dez anos na área de análise de dados e tive a sorte de trabalhar com muitas das pessoas de grande destaque na área. E uma das lições mais importantes que aprendi foi: uma boa ciência de dados é menos complicada do que as pessoas pensam. A melhor ciência de dados, na verdade, é surpreendentemente intuitiva.

O que torna a ciência de dados intuitiva? Na essência, ela trata de identificar padrões e prever o quanto uma variável afetará outra. As pessoas fazem isso o tempo todo.

Pense em como minha avó me deu um conselho amoroso. Ela utilizou uma grande base de dados de relacionamentos que seu cérebro coletou ao longo de quase um século de vida — das histórias que ouvia de familiares, amigos e conhecidos. Ela limitou sua análise a uma amostra de relacionamentos em que o homem tinha muitas de minhas características — um temperamento sensível, uma tendência a se isolar, senso de humor. Ela apontou as principais particularidades da mulher — gentileza, inteligência, beleza. E as correlacionou com as qualidades-chave de um bom relacionamento. Finalmente, informou os resultados. Em outras palavras, ela identificou padrões e previu o quanto uma variável afetaria outra. Vovó é uma cientista de dados.

E você também é. Quando era criança, você percebeu que quando chorava, sua mãe lhe dava atenção. Isto é ciência de dados. Quando chegou à fase adulta, notou que se você reclama demais, as pessoas evitam sua

companhia. Isto também é ciência de dados. Quando as pessoas de distanciam, você percebe que fica menos feliz. E quando está menos feliz, fica menos amigável. Quando fica menos amigável, as pessoas se afastam ainda mais. Ciência de dados. Ciência de dados. Ciência de dados.

Como a ciência de dados é tão natural, os melhores estudos de Big Data, eu descobri, podem ser entendidos por qualquer pessoa esperta. Se não for capaz de compreender um estudo, o problema não é você, mas o estudo.

Quer uma prova de que uma excelente ciência de dados tende a ser intuitiva? Recentemente, me deparei com um estudo que pode ser um dos mais importantes conduzidos nos últimos anos. Era ainda um dos mais intuitivos que já vi. Quero que você pense não apenas em sua importância — mas no quanto ele é natural e “estilo vovó”.

O estudo foi feito por uma equipe de pesquisadores da Microsoft e da Universidade de Columbia. A equipe queria descobrir quais sintomas previam o câncer de pâncreas. Esta doença tem uma taxa de sobrevivência de cinco anos — de apenas 3% —, mas a detecção precoce pode duplicar as chances do paciente.

Qual é o método dos pesquisadores? Eles utilizaram dados de dezenas de milhares de usuários anônimos do Bing, o mecanismo de busca da Microsoft. Codificaram um usuário como tendo recebido recentemente um diagnóstico de câncer de pâncreas com base em buscas inequívocas, tais como “acabei de ser diagnosticado com câncer de pâncreas” ou “descobri que tenho câncer de pâncreas, o que devo esperar”.

Em seguida, os pesquisadores analisaram as buscas por sintomas. Eles compararam o pequeno número de usuários que posteriormente reportaram um diagnóstico de câncer de pâncreas com os que não informaram a confirmação. Em outras palavras, quais sintomas previram que, dentro de algumas semanas ou meses, um usuário estaria relatando o diagnóstico?

Os resultados foram impressionantes. As buscas por dores nas costas e depois por amarelamento da pele mostraram ser sinais de câncer de pâncreas; as buscas apenas por dores nas costas indicaram a improbabilidade de alguém ter câncer de pâncreas. De modo semelhante, buscas por indigestão e depois dores abdominais revelaram-se prova de câncer pancreático, enquanto que buscas apenas por indigestão sem dores abdominais significaram que era improvável que a pessoa o tivesse. Esses pesquisadores conseguiram identificar de 5% a 15% dos casos com quase nenhum falso positivo. Pode não parecer uma taxa extraordinária, mas se você tem câncer de pâncreas, mesmo 10% de chances de dobrar suas possibilidades de sobrevivência são como ganhar na loteria.

O artigo detalhando este estudo seria difícil para não especialistas entenderem completamente. Inclui muito jargão técnico, como o teste Kolmogorov-Smirnov, cujo significado, devo admitir, tinha esquecido. (É um meio de determinar se um modelo se adéqua corretamente aos dados.)

Entretanto, observe como esse estudo notável é manifesto e intuitivo em seu nível mais elementar. Os pesquisadores analisaram uma ampla gama de casos médicos e tentaram conectar sintomas a doenças específicas. Você sabe quem mais usa esta metodologia para descobrir se uma pessoa tem ou não uma doença? Maridos e esposas, mães e pais, enfermeiras e médicos. Com base na experiência e no conhecimento, eles tentam conectar febres, dores de cabeça, corizas e dores de estômago a diversas doenças. Em outras palavras, os pesquisadores da Columbia e da Microsoft escreveram um estudo pioneiro utilizando a metodologia empírica e óbvia que todo mundo usa para fazer diagnósticos médicos.

Mas, espere. Vamos desacelerar um pouco. Se a metodologia da melhor ciência de dados é frequentemente natural e intuitiva, como alego, levanta uma questão fundamental sobre o valor do Big Data. Se seres humanos são naturalmente cientistas de dados, e a ciência de dados lhes é inerente, por que precisamos de computadores e softwares estatísticos? Por que precisamos do teste Kolmogorov-Smirnov? Não podemos apenas usar nossa intuição? Fazer como minha avó, as enfermeiras e os médicos?

Isto leva a uma discussão intensificada depois do lançamento do best-seller de Malcolm Gladwell, *Blink — A decisão num piscar de olhos*, que exalta a magia nas intuições das pessoas. Gladwell narra histórias de pessoas que, confiando unicamente em sua intuição, sabem dizer se uma estátua é falsa; se um jogador de tênis vai errar antes que acerte a bola; quanto um cliente está disposto a pagar. Os heróis em *Blink* não executam regressões; não calculam intervalos de confiança; não realizam testes Kolmogorov-Smirnov. Mas normalmente fazem previsões notáveis. Muitas pessoas apoiaram intuitivamente a defesa de Gladwell baseadas na intuição: confiam em suas impressões e sentimentos. Os fãs de *Blink* devem exaltar a sabedoria de minha avó ao dar conselhos amorosos sem a ajuda de computadores. Os fãs de *Blink* podem ser menos propensos a evocar meus estudos ou outros mostrados neste livro, que usam computadores. Se Big Data — do tipo computacional e não do tipo de minha avó — é uma revolução, ele precisa provar que é mais poderoso do que nossa intuição desassistida, que, como mostrado por Gladwell, é bastante impressionante.

O estudo da Columbia e da Microsoft oferece um exemplo claro de rigorosa ciência de dados e computadores nos ensinando coisas que apenas nossa intuição jamais será capaz de descobrir. Esse também é um caso em que o tamanho do conjunto de dados importa. Às vezes, há experiências insuficientes para que nossa intuição consiga analisar sozinha. É improvável que você — ou seus amigos próximos e familiares — tenha visto casos suficientes

de câncer de pâncreas para descobrir a diferença ente indigestão seguida de dor abdominal e indigestão isolada. De fato, é inevitável que, à medida que os conjuntos de dados do Bing ficam maiores, os pesquisadores identifiquem muitos outros padrões sutis na sequência de sintomas — para esta e outras doenças — que até os médicos ignoraram.

Além do mais, enquanto nossa intuição normalmente nos dá um bom senso de como o mundo funciona, em geral, não é exata. Precisamos de dados para avivar a imagem. Considere, por exemplo, os efeitos do clima no humor. Você já deve imaginar que pessoas são mais propensas a se sentir deprimidas em um dia de -12°C do que em um de 22°C. De fato, isso é verdade. Mas você provavelmente não imagina qual é o tamanho do impacto dessa diferença de temperaturas. Pesquisei pelas correlações entre as buscas no Google por depressão e uma ampla gama de fatores, incluindo condições econômicas, níveis de educação e frequência a igrejas em determinada região. O clima de inverno superou todas. Em meses de inverno nos Estados Unidos, regiões mais amenas, como Honolulu, no Havaí, têm 40% menos buscas por depressão do que lugares frios, como Chicago, Illinois. Mas quanto desse efeito é significativo? Uma leitura otimista sobre a eficácia dos antidepressivos concluiria que as drogas mais eficazes diminuem a incidência de depressão em cerca de 20%. A julgar pelos números do Google, uma mudança de Chicago para Honolulu seria pelo menos duas vezes mais eficaz do que os medicamentos para sua tristeza de inverno.*

Às vezes nossa intuição, quando não orientada por uma análise computacional cuidadosa, está redondamente enganada. Podemos ser ofuscados por experiências e preconceitos pessoais. Na verdade, apesar de minha avó ter sido capaz de utilizar suas décadas de experiência para oferecer melhores conselhos amorosos do que o resto de minha família, ainda tinha algumas visões duvidosas sobre o que faz um relacionamento ser duradouro. Por exemplo, ela frequentemente enfatiza a importância de ter amigos em comum. Acredita que isso foi um fator-chave no sucesso de seu casamento: ela passava quase todas as noites amenas com seu marido, meu avô, em seu pequeno quintal no Queens, Nova York, sentada em espreguiçadeiras fofocando com um grupo de vizinhos.

Entretanto, arriscando-me a arruinar a carreira de casamenteira de minha avó, a ciência de dados sugere que sua teoria está errada. Uma equipe de cientistas de computação recentemente analisou o maior conjunto de dados jamais coletado sobre relacionamento humano — o Facebook. Estudaram um grande número de casais que estavam em determinado momento “em um relacionamento sério”. Alguns desses casais permaneceram “em um relacionamento sério”. Outros mudaram seus status para “solteiro(a)”. Ter um grupo de amigos em comum, os pesquisadores descobriram, é um forte preditor de que o relacionamento *não* irá durar. Talvez sair toda noite com seu parceiro e o mesmo pequeno grupo de amigos não seja uma coisa boa; círculos sociais separados podem ajudar a criar um relacionamento mais sólido.

Como você pode ver, apenas nossa intuição, sem a ajuda de computadores, às vezes impressiona. Mas também comete grandes equívocos. Vovó pode ter caído na armadilha cognitiva: a propensão para exagerar a importância da própria experiência. Na linguagem dos cientistas de dados, *ponderamos* nossos dados e tendemos a atribuir muito mais peso a um dado em particular: nós mesmos.

Minha avó estava tão focada em suas fofocas noturnas com meu avô e os amigos deles que não pensou o bastante sobre outros casais. Ela se esqueceu de considerar seu cunhado e a esposa, que papeavam quase toda noite com o mesmo pequeno grupo de amigos, mas que brigavam o tempo todo até se divorciarem. Ignorou totalmente meus pais, a filha e o genro dela. Meus pais saíam separados muitas vezes — meu pai ia a clubes de jazz ou a jogos esportivos, minha mãe, para um restaurante ou o teatro com suas amigas; e ainda assim permaneceram casados e felizes.

Ao confiar em nossa intuição, também somos atraídos pelo fascínio humano elementar pelo drama. Tendemos a superestimar a prevalência de tudo que torna uma história memorável. Por exemplo, quando questionadas em uma pesquisa, as pessoas reiteradamente classificam tornados como uma causa mais comum de morte do que a asma. Na verdade, a asma provoca aproximadamente 70 vezes mais mortes. No entanto, mortes por asma não se sobressaem — e não aparecem nos noticiários. As mortes por tornados sim.

Em outras palavras, costumamos nos equivocar sobre como o mundo funciona quando confiamos apenas no que ouvimos ou em nossa experiência pessoal. Enquanto a metodologia de uma boa ciência de dados é geralmente intuitiva, os resultados são frequentemente contraintuitivos. A ciência de dados pega um processo humano nacional e manifesto — a identificação de padrões e a tentativa de os decifrar — e injeta esteroides, potencialmente nos mostrando que o mundo funciona de uma maneira completamente diferente do que pensávamos. Foi isso que aconteceu quando estudei os preditores de sucesso no basquete.

Quando eu era criança, tinha um sonho, e um sonho apenas: queria ser economista ou cientista de dados quando crescesse. Não. Estou brincando. Queria desesperadamente ser um jogador de basquete profissional, para seguir os

passos de meu herói, Patrick Ewing, o maior pivô da história do *New York Knicks*.

Às vezes, chego a pensar que dentro de cada cientista de dados existe um garoto tentando descobrir por que seus sonhos de infância não se tornaram realidade. Assim, não é de surpreender que recentemente eu tenha investigado o que é preciso para chegar à NBA. Os resultados da investigação foram inusitados. Na verdade, demonstraram mais uma vez o quanto uma boa ciência de dados muda sua visão do mundo e o quanto contraintuitivos os números são.

A questão específica que investiguei foi: você tem mais probabilidade de chegar à NBA se crescer em uma família pobre ou de classe média?

Muitas pessoas diriam: pobre. A sabedoria convencional diz que crescer em circunstâncias difíceis, talvez em condomínios para pessoas de baixa renda com uma mãe solteira e adolescente, ajude a fomentar a determinação necessária para alcançar os níveis mais altos desse esporte intensamente competitivo.

Esta nova visão foi expressa por William Ellerbee, treinador de basquete de uma escola de ensino médio na Filadélfia, em uma entrevista para a *Sports Illustrated*. "Garotos de bairros de classe média costumam jogar pela diversão", disse Ellerbee. "Garotos de áreas mais pobres veem o basquete como questão de vida e morte." Eu, infelizmente, fui criado por pais casados, em um bairro de classe média de Nova Jersey. LeBron James, o melhor jogador de minha geração, nasceu pobre, filho de uma jovem solteira de 16 anos em Akron, Ohio.

De fato, uma pesquisa na internet que realizei sugeriu que a maioria dos norte-americanos pensa o mesmo que o técnico Ellerbee e eu pensávamos: que a maior parte dos jogadores da NBA cresceu na pobreza.

A sabedoria convencional está correta?

Vamos analisar os dados. Não há fontes de dados abrangentes sobre as condições socioeconômicas dos jogadores da NBA. Mas sendo detetives de dados, utilizando uma porção de fontes — basketball-reference.com, ancestry.com, o censo dos Estados Unidos e outras —, descobrimos qual contexto familiar realmente conduz mais jogadores à NBA. Este estudo, você perceberá, utiliza uma variedade de fontes de dados, algumas maiores, outras menores, algumas online, outras offline. Por mais empolgantes que algumas fontes sejam, um bom cientista de dados não despreza as tradicionais que podem ajudar. A melhor maneira de obter a resposta correta para uma pergunta é combinar todos os dados disponíveis.

O primeiro dado relevante é o local de nascimento de cada jogador. Para todos os condados nos Estados Unidos, anotei quantos homens negros e brancos nasceram na década de 1980. Depois, registrei quantos chegaram à NBA. Comparei essa informação com a média de rendimentos domiciliar desse condado em particular. Também fiz controles demográficos raciais do condado, pois — e esse é um assunto para um livro inteiro — homens negros têm 40% mais probabilidade de chegar à NBA do que os brancos.

Os dados nos dizem que um homem tem uma chance substancialmente melhor de chegar à NBA se nascer em um condado rico. Um garoto negro nascido em um dos condados mais ricos nos Estados Unidos, por exemplo, tem duas vezes mais chances de chegar à NBA do que um garoto negro nascido nos condados mais pobres. Para um garoto branco, a vantagem de ter nascido em um dos condados mais ricos comparada a nascer em um dos mais pobres é de 60%.

Isto sugere, contrariando a sabedoria convencional, que homens pobres são, na verdade, sub-representados na NBA. Entretanto, estes dados não são perfeitos, pois muitos condados ricos nos Estados Unidos, como o Condado de Nova York (Manhattan), também incluem bairros pobres, como o Harlem. Então, ainda é possível que uma infância difícil ajude um jogador a chegar à NBA. Ainda precisamos de mais pistas, mais dados.

Assim, investiguei contextos familiares de jogadores da NBA. As informações foram encontradas em histórias de noticiários e em redes sociais. Esta metodologia demandou muito tempo, então limitei a análise aos cem jogadores da NBA afro-americanos nascidos na década de 1980 que marcaram mais pontos. Comparados ao homem negro médio nos Estados Unidos, os superastros da NBA tinham 30% menos probabilidade de terem nascido de mães adolescentes ou solteiras. Em outras palavras, o contexto familiar dos melhores jogadores negros da NBA ainda sugere que um ambiente confortável é uma grande vantagem para alcançar o sucesso.

Dessa forma, nem os dados do nível do condado de nascimento nem o contexto familiar de uma amostra limitada de jogadores nos oferece uma informação perfeita sobre as infâncias de todos os jogadores da NBA. Assim, não estou inteiramente convencido de que famílias com pais casados, de classe média, produzam mais estrelas da NBA do que famílias pobres e mães solteiras. Quanto mais dados pudermos trazer para esta questão, melhor.

Então, me lembrei de mais um ponto de dado que oferece pistas promissoras sobre o histórico de um homem. Existe um artigo escrito por dois economistas, Roland Fryer e Steven Levitt, que sugere que o primeiro nome de uma pessoa negra é um indicador de seu contexto socioeconômico. Fryer e Levitt estudaram certidões de nascimento na Califórnia da década de 1980, e descobriram que, entre os afro-americanos, mães adolescentes, solteiras, pobres e de baixa instrução tendem a dar a seus filhos nomes diferentes do que pais casados, de classe média e instruídos.

Filhos de contextos sociais melhores têm maior probabilidade de receber nomes comuns, como Kevin, Chris e John. Filhos de lares em dificuldades, em condomínios de pessoas de baixa renda, têm maior probabilidade de receber nomes únicos, como Knowshon, Uneek e Breionshay. Crianças afro-americanas nascidas na pobreza têm quase duas vezes mais chances de receber um nome diferente de todas as outras nascidas naquele mesmo ano.

E quanto aos nomes dos jogadores negros da NBA? Parecem mais de negros pobres ou de classe média? Observando-se o mesmo período, jogadores da NBA nascidos na Califórnia têm metade da probabilidade de ter nomes únicos do que o homem negro médio, uma diferença estatisticamente significativa.

Conhece alguém que pensa que a NBA é uma liga para crianças do gueto? Diga para ouvir atentamente o próximo jogo. Peça que anote quantas vezes Russell dribla Dwight, depois tenta passar a bola para Josh e então para as mãos de Kevin. Se a NBA fosse realmente uma liga repleta de homens negros e pobres, a narração seria bem diferente. Haveria muito mais homens com nomes como LeBron.

Agora, reunimos três diferentes conjuntos de provas — o condado de nascimento, o estado civil das mães dos maiores pontuadores e o primeiro nome dos jogadores. Nenhuma fonte é perfeita. Mas todas embasam a mesma história. Melhor status socioeconômico significa chances maiores de entrar na NBA. Ou seja, a sabedoria convencional está errada.

Dentre todos os afro-americanos nascidos na década de 1980, cerca de 60% tinha pais não casados. Mas estimo que entre os afro-americanos nascidos na mesma década que chegaram à NBA, uma significativa maioria tinha pais casados. Em outras palavras, a NBA não é composta principalmente de homens com históricos como o de LeBron James. Há mais homens como Chris Bosh, criado pelos pais, no Texas, que incentivaram seu interesse por eletrônicos, ou Chris Paul, o segundo filho de pais de classe média de Lewisville, Carolina do Norte, cuja família participou com ele de um episódio de *Family Feud* em 2011.

O objetivo de um cientista de dados é compreender o mundo. Uma vez que encontramos um resultado contraintuitivo, podemos usar mais ciência de dados para nos explicar por que o mundo não é o que parece. Por que, por exemplo, homens de classe média têm uma vantagem no basquete em relação aos pobres? Há pelo menos duas explicações.

Primeira, porque homens pobres tendem a ficar mais baixos. Acadêmicos sabem há bastante tempo que cuidados de saúde e nutrição na infância desempenham um papel importante na saúde dos adultos. É por isso que o homem médio no mundo desenvolvido é hoje dez centímetros mais alto do que há um século e meio. Os dados sugerem que norte-americanos de comunidades subdesenvolvidas, devido a nutrição e cuidados de saúde mais deficientes na infância, são mais baixos.

Os dados também nos mostram o efeito da altura para chegar à NBA. Sem dúvidas você intuiu que ser alto é de grande ajuda quando se pretende ser jogador de basquete. Basta comparar a altura de um típico jogador na quadra em relação a um fã na arquibancada. (A altura média de um jogador da NBA é 2,04m; a altura média de um norte-americano é 1,79m.)

Quanto a altura é importante? Os jogadores da NBA às vezes mentem sobre suas alturas e não existe uma listagem completa sobre a distribuição de altura dos homens norte-americanos. Mas trabalhando com uma estimativa matemática simples de como seria esta distribuição de altura e os números da própria NBA, é fácil confirmar que os efeitos da altura são gigantescos — talvez ainda maiores do que podemos ter suspeitado. Estimo que cada 2,5 centímetros adicionais duplique as chances de se chegar à NBA. E isso é verdade para toda a extensão de distribuição de altura. Um homem de 1,80m tinha duas vezes mais chances de chegar à NBA do que um de 1,78m. Um homem de 2,10m, duas vezes mais chances do que um de 2,07m. Ao que parece, entre homens abaixo de 1,83m de altura, apenas cerca de um a cada dois milhões chega à NBA. Dentre os acima de 2,13m, eu e outros analistas estimamos que algo em torno de um a cada cinco chegue à NBA.

Os dados, você verá, esclarecem por que meu sonho de estrelato no basquete descarrilou. Não foi porque cresci em um bairro de classe média. E sim porque tenho 1,75m e sou branco (sem mencionar lento). Sou preguiçoso, também. E não tenho muita resistência, meu arremesso é terrível e ocasionalmente sofro ataques de pânico quando a bola está em minhas mãos.

A segunda razão pela qual garotos com históricos familiares complicados podem ter dificuldade de chegar à NBA é que às vezes carecem de certas habilidades sociais. Usando dados em milhares de crianças em idade escolar, os economistas descobriram que famílias de classe média, com ambos os pais, são em média substancialmente melhores na criação de filhos confiantes, disciplinados, persistentes, focados e organizados.

Então, como a pouca habilidade social pode arruinar uma carreira promissora no basquete?

Vamos analisar a história de Doug Wrenn, uma das esperanças mais talentosas do basquete nos anos 1990. Seu técnico universitário, Jim Calhoun, na Universidade de Connecticut, que treinou futuros astros da NBA, afirmava que Wrenn pulava mais alto do que qualquer um com quem tinha trabalhado. Mas Wrenn teve uma criação desafiadora. Foi criado pela mãe solteira em Blood Alley, um dos bairros mais violentos de Seattle. Em

Connecticut, ele vivia em conflito com todos ao redor. Insultava jogadores, questionava técnicos e usava roupas largas, violando as regras do time. Ele também tinha problemas com a lei — roubou sapatos de uma loja e insultou policiais. Calhoun finalmente se cansou e o expulsou do time.

Wrenn recebeu uma segunda chance, na Universidade de Washington. Mas lá, também, uma inabilidade de conviver com outras pessoas arruinou sua carreira. Ele brigou com o treinador por causa de tempo de jogo e escolha de jogadas, e foi expulso do time. Wrenn foi desconvocado pela NBA, tentou algumas ligas menores, e foi morar com a mãe até ser preso por agressão. "Minha carreira acabou", disse Wrenn para o *Seattle Times* em 2009. "Meus sonhos, minhas aspirações acabaram. Doug Wrenn está morto. Aquele jogador de basquete, aquele cara morreu. Acabou." Wrenn tinha talento não apenas para ser um jogador da NBA, mas para ser um jogador fantástico, lendário. Porém nunca desenvolveu o temperamento para sequer se manter em um time universitário. Talvez se tivesse tido uma vida mais estável na infância, pudesse ter sido o próximo Michael Jordan.

Michael Jordan, é claro, também tinha um impressionante salto vertical. Além de um grande ego e uma competitividade feroz — uma personalidade às vezes não muito diferente da de Wrenn. Jordan poderia ter sido um garoto difícil. Aos 12 anos, foi expulso da escola por brigar. Mas tinha ao menos uma coisa que Wrenn não tinha: uma criação estável de classe média. Seu pai era supervisor de equipamento da General Electric; sua mãe, bancária. E eles o ajudaram a administrar sua carreira.

Na verdade, a vida de Jordan está repleta de histórias de sua família o desviando das armadilhas em que um grande e competitivo talento pode acabar se metendo. Depois que Jordan foi expulso da escola, sua mãe passou a levá-lo junto para o trabalho. Ele não podia sair do carro e tinha que permanecer sentado no estacionamento lendo livros. Depois que foi recrutado pelo Chicago Bulls, seus pais e irmãos revezavam as visitas para se certificar de que ele não caísse nas tentações que a fama e o dinheiro trazem.

A carreira de Jordan não terminou como a de Wrenn, com uma citação sem muita importância no *Seattle Times*. Terminou com um grande discurso na cerimônia de admissão no Hall da Fama do Basquete, assistido por milhões de pessoas. Em seu discurso, Jordan disse que tentou manter o "foco nas coisas boas da vida — saber como as pessoas o veem, como você as respeita… como é percebido publicamente. Parar um momento e pensar sobre as coisas que faz. E tudo isso veio dos meus pais".

Os dados nos dizem que Jordan está absolutamente certo em agradecer a seus pais casados de classe média. Os dados indicam que em famílias e comunidades mais pobres, existem talentos no nível da NBA que não estão na NBA. Esses homens tinham os genes e a ambição, mas nunca desenvolveram o temperamento para se tornar superastros do basquete.

E não — seja lá o que possamos ter intuído — estar em circunstâncias tão desesperadoras que o basquete pareça uma "questão de vida ou morte" não ajuda. Histórias como a de Doug Wrenn ilustram isso. E os dados comprovam.

Em junho de 2013, LeBron James foi entrevistado na televisão depois de vencer seu segundo campeonato da NBA. (Desde então, já venceu um terceiro.) "Eu sou LeBron James", declarou ele. "De Akron, Ohio. Do gueto. Eu não devia nem estar aqui." O Twitter e outras redes sociais explodiram de críticas. Como uma pessoa incrivelmente talentosa, reconhecida ainda muito jovem como o futuro do basquete, alega ser um coitado? Na verdade, qualquer um vindo de um ambiente difícil, não importa suas proezas atléticas, tem probabilidades desfavoráveis. Em outras palavras, as realizações de James são ainda mais excepcionais do que parecem a princípio. Os dados também provam isso.

* Revelação completa: Logo após concluir este estudo, mudei da Califórnia para Nova York. Usar os dados para saber o que você deve fazer é geralmente fácil. O difícil mesmo é colocar em prática.

PARTE II

OS PODERES DO BIG DATA

2

FREUD ESTAVA CERTO?

Recentemente vi uma pessoa andando pela rua ser descrita como “penistrian”. Entendeu? “Penistrian” em vez de “pedestrian” [pedestre]. Essa nova palavra surgiu em um grande conjunto de dados de erros de digitação que as pessoas cometem. Uma pessoa vê alguém caminhando e escreve a palavra “penis”. Isso deve significar alguma coisa, não é mesmo?

Recentemente soube de um homem que sonhou que comia uma banana enquanto caminhava até o altar para se casar com sua esposa. Descobri essa informação em um grande conjunto de dados de sonhos que as pessoas gravam em um aplicativo. Um homem imagina estar se casando com uma mulher enquanto come um alimento em formato fálico. Isso também deve significar alguma coisa, não é mesmo?

Será que Sigmund Freud estava certo? Desde que suas teorias chegaram a público pela primeira vez, a resposta mais honesta para esta pergunta seria dar de ombros. Foi Karl Popper, filósofo austríaco-britânico, que esclareceu mais esta questão. Popper notoriamente declarou que as teorias de Freud não podiam ser refutadas. Não havia um modo de testar se eram verdadeiras ou falsas.

Freud poderia dizer que a pessoa que escreveu “penistrian” revela um possível desejo sexual reprimido. A pessoa poderia responder que não estava revelando nada; que poderia da mesma forma ter cometido outro erro de digitação inocente, como “pedaltrian”. Cada um teria uma versão e nenhuma prova. Freud poderia dizer que o cavalheiro que sonhou que comia uma banana no dia de seu casamento secretamente pensava em um pênis, revelando seu desejo de preferir se casar com um homem. O cavalheiro por sua vez diria que apenas sonhou com um banana. Ele poderia ter sonhado que comia uma maçã enquanto caminhava até o altar. Seria um outro impasse. Não havia uma forma de realmente testar a teoria de Freud.

Quer dizer, até agora.

A ciência de dados torna muitas concepções de Freud refutáveis — ela coloca suas famosas teorias em xeque. Vamos começar com símbolos fálicos em sonhos. Usando um gigantesco conjunto de dados de sonhos registrados, podemos facilmente perceber com que frequência objetos de formato fálico aparecem. Comida é um bom lugar para focar este estudo. Alimentos aparecem em muitos sonhos, e muitos têm formato fálico — bananas, pepinos, cachorros-quentes etc. Podemos então mensurar os fatores que nos fazem sonhar mais com certos alimentos do que com outros — com que frequência as pessoas os comem, o quanto gostam deles e, sim, se são fálicos em essência.

Podemos testar se duas comidas, igualmente populares, mas apenas com formato fálico, aparecem em sonhos em quantidades diferentes. Se os alimentos em forma fálica não apresentarem maior probabilidade de aparecer em sonhos, então símbolos fálicos não são fatores significativos nos sonhos. Graças ao Big Data, esta parte da teoria de Freud pode ser, de fato, refutável.

Recebi dados do Shadow, um aplicativo que pede ao usuário que registre seus sonhos. Codifiquei os alimentos listados em dezenas de milhares de sonhos.

De modo geral, o que nos faz sonhar com comida? O principal preditor é a frequência em que a consumimos. A substância que mais aparece em sonhos é a água. Os vinte alimentos mais comuns incluem frango, pão, sanduíches e arroz — todos particularmente não freudianos.

O segundo preditor da frequência com que um alimento aparece em sonhos é o quanto a pessoa o acha saboroso. As duas comidas com que sonhamos com mais frequência são notadamente não freudianas, mas excepcionalmente saborosas: chocolate e pizza.

Então, e quanto aos alimentos com formato fálico? Eles invadem nossos sonhos com constância inesperada? Não.

Bananas são a segunda fruta mais comum a aparecer em sonhos. Mas são também a segunda mais consumida. Então não precisamos de Freud para explicar a frequência com que sonhamos com bananas. Pepinos são o sétimo vegetal mais comum a aparecer em sonhos. Eles são o sétimo vegetal mais consumido. Então, de novo, seu formato não é necessário para explicar sua presença em nossas mentes enquanto dormimos. Sonhos com cachorros-quentes são muito menos comuns do que com hambúrgueres. Isso é verdadeiro até mesmo pelo padrão de comparação de que as pessoas comem mais hambúrgueres do que cachorros-quentes.

Em geral, usando a análise de regressão (um método que permite aos cientistas sociais identificar separadamente o impacto de múltiplos fatores) em todas as frutas e vegetais, descobri que o fato de uma comida ter formato fálico não resulta em maior probabilidade de aparecer em sonhos do que seria esperado pela sua popularidade. Esta teoria de Freud é refutável — e, pelo menos de acordo com a minha visão dos dados, falsa.

Em seguida, considere o lapso freudiano, ou ato falho. O psicólogo teorizou que nossos erros — as maneiras que erramos ao dizer ou escrever algo — revelam desejos subconscientes, frequentemente sexuais. Podemos usar Big Data para testar esta hipótese? Existe uma maneira: ver se nossos erros — nossos lapsos — tendem a ser maliciosos. Se nossos desejos sexuais ocultos se mostram neles, deve haver um número desproporcional de erros que incluam palavras como "penis" [pênis], "cock" [pinto] e "sex" [sexo].

Por isso estudei um conjunto de dados de mais de 40 mil erros de digitação coletados por pesquisadores da Microsoft. O conjunto de dados incluía erros cometidos e imediatamente corrigidos. Dentre essas dezenas de milhares de erros, havia muitos indivíduos cometendo deslizes de natureza sexual. Havia o já mencionado "penistrian"; alguém que digitou "*sex*urity" em vez de "security" [segurança] e "cocks" [pintos] em vez de "rocks" [pedras]. Mas também havia muitos lapsos inocentes. As pessoas escreveram "pindows" (windows = janelas ou o software) e "fegetables" (vegetables = vegetais), "aftermoons" (afternoons = tardes) e "refriderators" (refrigerators = geladeiras).

Então, o número de lapsos sexuais foi destoante?

Para testar isso, primeiro usei o conjunto de dados da Microsoft para modelar com que frequência as pessoas trocam letras específicas. Calculei quantas vezes substituíam um *t* por um *s*, um *g* por um *h*. Em seguida, criei um programa de computador que comete os erros do mesmo modo que as pessoas. Podemos chamá-lo de *Error Bot*. O programa substituiu um *t* por um *s* com a mesma frequência que os seres humanos no estudo da Microsoft. Fez o mesmo com *g* por um *h*. E assim por diante. Executei o programa usando as mesmas palavras que as pessoas do estudo da Microsoft escreveram errado. Em outras palavras, Error Bot tentou escrever "pedestrian", "rocks", "windows" e "refrigerator." Mas substituiu um *r* por um *t* com a mesma frequência que as pessoas, e escreveu, por exemplo, "tocks". Substitui um *r* por um *c* com a mesma frequência que os seres humanos, e escreveu "cocks".

Então, o que aprendemos com a comparação entre Error Bot e pessoas normalmente desatentas? Depois de cometer alguns milhões de erros, apenas substituindo as letras das mesmas maneiras que os seres humanos, Error Bot incorreu em numerosos erros de natureza freudiana. Escreveu "seashell" como "*sex*shell", "lipstick" como "lips*dick*" e "luckiest" como "*fuck*iest", juntamente com muitos outros erros similares. E — este é o ponto principal — Error Bot, que é claro que não tem subconsciente, apresentou a mesma probabilidade de cometer erros de natureza sexual que pessoas. Com a ressalva, como nós, cientistas sociais, gostamos de dizer, que é preciso haver mais pesquisa, isto significa que erros de conotação sexual não têm maior probabilidade de ser cometidos por seres humanos do que o esperado pelo mero acaso.

Isto quer dizer que para que as pessoas cometam erros de digitação como "penistrian", "sexurity" e "cocks" não é preciso que haja uma conexão entre o erro e o proibido, alguma teoria da mente em que as pessoas revelem seus desejos secretos por meio de seus erros. Esses tipos de lapsos são explicados inteiramente pela frequência normal dos erros de digitação. As pessoas cometem muitos erros. E se forem suficientes, eventualmente serão coisas como "lipsdick", "fuckiest" e "penistrian". Se um macaco digitar por tempo bastante, acabará escrevendo "ser ou não ser". Se uma pessoa digitar por tempo suficiente, em algum momento escreverá "penistrian".

A teoria de Freud de que os erros revelam os desejos de nosso subconsciente é, de fato, refutável — e, de acordo com minha análise dos dados, falsa.

Big Data nos diz que uma banana é sempre somente uma banana, e que um "penistrian" é apenas "pedestrian" grafado errado.

Então, Freud estava totalmente enganado em todas as teorias? Não exatamente. A primeira vez que obtive acesso aos dados do PornHub, descobri uma revelação que me pareceu pelo menos um tanto freudiana. Na verdade, ela está entre as coisas mais surpreendentes que encontrei durante minhas investigações sobre os dados: um número chocante de pessoas que visitam sites pornô populares busca imagens de incesto.

Dentre as cem buscas mais comuns no PornHub, um dos sites pornô mais populares, dezesseis procuram por vídeos de temática incestuosa. Aviso prudente — serei um tanto explícito: incluíam “irmão e irmã”, “madrasta fodendo enteado”, “mãe e filho”, “mãe fode filho” e “irmão e irmã de verdade”. A pluralidade de buscas incestuosas masculinas é por cenas envolvendo mães e filhos. E as mulheres? Nove das cem buscas mais comuns feitas por mulheres no PornHub são por vídeo com temática incestuosa e contendo fantasias semelhantes — mas com os gêneros dos pais ou dos filhos envolvidos normalmente invertidos. Assim, a pluralidade de buscas incestuosas feitas por mulheres é por cenas com pais e filhas.

Não é difícil identificar nestes dados pelo menos um suave eco do Complexo de Édipo de Freud. Ele teorizava a existência de um desejo quase universal na infância, que mais tarde é reprimido, de envolvimento sexual pelo genitor do sexo oposto. Quem dera que o psicólogo vienense tivesse vivido o bastante para aplicar suas habilidades analíticas aos dados do PornHub, em que o interesse no genitor do sexo oposto parece ser mantido por adultos — de modo bastante explícito — e pouco é reprimido.

Obviamente, os dados do PornHub não podem nos dizer com certeza sobre o que as pessoas estão fantasiando ao assistir a esses vídeos. Será que estão mesmo imaginando fazer sexo com os próprios genitores? As buscas no Google podem nos dar mais algumas pistas de que existem muitas com desejos deste tipo.

Considere todas as buscas na forma “Eu quero fazer sexo com meu/minha…” A maneira mais comum de completar esta busca é “mãe”. Em geral, mais de três quartos das buscas deste formato são incestuosas. E isso não se deve à construção específica da frase. Buscas na forma “Estou atraído por”, por exemplo, têm ainda mais prevalências de admissões de desejos incestuosos. Mas devo dizer — arriscando-me a decepcionar *Herr* Freud — que essas não são buscas especialmente comuns: alguns poucos milhares de pessoas a cada ano em todos os Estados Unidos admitem ter atração pela própria mãe. Alguém também deveria contar para Freud que as buscas no Google, como discuto mais adiante neste livro, às vezes tendem na direção do proibido.

No entanto, as pessoas experimentam um grande número de outras atrações inapropriadas que eu esperava que tivessem sido mencionadas com mais frequência nas buscas. Chefe? Empregado? Aluno? Terapeuta? Paciente? Melhor amiga da esposa? Filha do melhor amigo? Irmã da esposa? Esposa do melhor amigo? Nenhum destes desejos confessos podem concorrer com a mãe. Talvez, se combinarmos estes aos dados do PornHub, isso realmente possa significar alguma coisa.

E a asserção geral de Freud de que a sexualidade é moldada pelas experiências de infância é amparada por outro aspecto nos dados do Google e do PornHub, que revelam que homens, pelo menos, retêm um excessivo número de fantasias relacionadas à infância. De acordo com as buscas de esposas sobre os maridos, alguns dos fetiches mais comuns de homens adultos são o desejo de usar fraldas e ser amamentado, especialmente, como mencionei antes, na Índia. Além do mais, desenhos animados pornôs — animação de cenas de sexo explícito com personagens de programas populares entre garotos adolescentes — alcançaram um alto nível de popularidade. Ou considere as ocupações de mulheres mais buscadas por homens em sites pornô. Homens entre 18 e 24 anos buscam com mais frequência por babás. Assim como os homens entre 25 e 64 anos. E homens a partir de 65 anos. E para os homens em todos os grupos etários, professoras e líderes de torcida estão entre as quatros mais populares. Claramente, os primeiros anos de vida parecem exercer um papel descomunal nas fantasias de homens adultos.

Ainda não fui capaz de usar todos estes dados inéditos sobre a sexualidade adulta para desvendar com precisão como se formam as preferências sexuais. Ao longo das próximas décadas, outros cientistas sociais e eu seremos capazes de criar novas teorias refutáveis sobre a sexualidade adulta e testá-las com dados reais.

Já consigo prever alguns temas básicos que certamente farão parte da teoria da sexualidade adulta baseada em dados. Claramente, não será uma história idêntica à contada por Freud, com seus estágios universais de infância e repressão específicos e bem definidos. Mas, com base em minha primeira análise nos dados do PornHub, estou absolutamente certo de que o veredito sobre a sexualidade adulta retratará alguns temas-chave enfatizados por Freud. A infância desempenhará um papel enorme. Assim como as mães.

Provavelmente, teria sido impossível analisar Freud dessa maneira dez anos atrás. Por certo, teria sido impossível há 80 anos, quando Freud ainda estava vivo. Assim, vamos pensar por que essas fontes de dados podem ajudar. Esse exercício ajuda a entender por que o Big Data é tão poderoso.

Lembre-se, dissemos que apenas ter montanhas e montanhas de dados por si só não gera automaticamente novas percepções. O tamanho dos dados, sozinho, é superestimado. Por que, então, o Big Data é tão poderoso? Por que

revolucionará o modo como vemos a nós mesmo? Considero que existem quatro poderes únicos do Big Data. Esta análise de Freud oferece uma boa ilustração deles.

Você pode ter percebido, em primeiro lugar, que estamos levando a pornografia a sério na discussão de Freud. E utilizamos com frequência dados de pornografia neste livro. De modo um tanto surpreendente, estes dados são raramente utilizados por sociólogos, muitos dos quais preferem confiar nas pesquisas em conjuntos de dados tradicionais sobre as quais construíram suas carreiras. Mas uma pequena reflexão mostra que o uso difundido de pornografia — e os dados de busca e de visualização decorrentes — é o modo mais importante de desenvolver nossa capacidade de entender a sexualidade humana em, bem… Na verdade, provavelmente é o mais importante até hoje. Esses dados fariam Schopenhauer, Nietzsche, Freud e Foucault salivar. Esses dados não existiam na época deles. Não existiam há algumas décadas. Existem hoje. Existem muitas fontes de dados únicas, sobre uma gama de tópicos, que nos oferece perspectivas sobre áreas que antes só podíamos adivinhar. *Oferecer novos tipos de dados é o primeiro poder do Big Data.*

Os dados de pornografia e os de busca no Google não são apenas novos; são honestos. Na era pré-digital, as pessoas escondiam seus pensamentos embaraçosos das outras pessoas. Na digital, ainda os escondem dos outros, mas não da internet ou em sites específicos, como Google e PornHub, que protegem seu anonimato. Esses sites funcionam como uma espécie de soro da verdade digital — por isso temos a habilidade de revelar o fascínio disseminado pelo incesto. Big Data nos permite finalmente ver o que as pessoas realmente querem e realmente fazem, não o que dizem que querem e que fazem. *Oferecer dados honestos é o segundo poder do Big Data.*

Como hoje existem tantos dados, há informação significativa em cada pequena porção de uma população. Podemos comparar, digamos, o número de pessoas que sonham com pepinos versus aquelas que sonham com tomates. *Permitir a análise de pequenos subconjuntos de pessoas é o terceiro poder do Big Data.*

O Big Data tem ainda mais um impressionante poder — um que não foi utilizado em meu rápido estudo de Freud, mas pode ser em um próximo: permite realizar experimentos rápidos e controlados. Isso permite testar a causalidade, não apenas as correlações. Esses tipos de testes são mais utilizados em empresas, hoje, mas se mostrarão uma poderosa ferramenta para os cientistas sociais. *Permitir a realização de muitos experimentos causais é o quarto poder do Big Data.*

É chegada a hora de detalhar esses poderes e explorar exatamente por que o Big Data é tão importante.

3

OS DADOS REINVENTADOS

Às 6 horas da manhã de uma determinada sexta-feira de cada mês, as ruas de grande parte de Manhattan estarão praticamente desertas. As lojas nessas ruas estarão fechadas, suas fachadas, cobertas por grades de ferro, os apartamentos, escuros e silenciosos.

Os pisos do Goldman Sachs, o banco de investimentos global na parte baixa de Manhattan, por outro lado, estarão intensamente iluminados, seus elevadores transportando milhares de trabalhadores até suas mesas. Às 7 horas da manhã a maioria das mesas estará ocupada.

Não seria injusto descrever esse momento em qualquer outro dia nesta parte da cidade como adormecido. Nesta manhã de sexta, no entanto, haverá um frenesi de energia e empolgação. Neste dia, a informação que impactará enormemente o mercado de ações está programada para chegar.

Minutos depois de sua divulgação, esta informação será repercutida nos sites de notícias. Segundos depois da divulgação, a informação será discutida, debatida e dissecada, ruidosamente, no Goldman e em centenas de outras instituições financeiras. Mas muito da ação real em finanças hoje em dia acontece em milésimos de segundos. O Goldman e outras instituições financeiras pagam dezenas de milhões de dólares para ter acesso a cabos de fibras óticas que reduzem o tempo em que a informação viaja de Chicago a Nova Jersey em apenas quatro milissegundos (de 17 para 13). Instituições financeiras têm algoritmos operando para ler a informação e negociar com base nela — tudo em questão de milissegundos. Depois que essa informação crucial é divulgada, o mercado se movimentará em menos tempo do que um piscar de olhos.

Mas que informação vital é essa tão valiosa para o Goldman e diversas outras instituições financeiras?

A taxa mensal de desemprego.

A taxa, porém — que tem tamanho impacto no mercado de ações a ponto de as instituições financeiras fazerem o que for necessário para maximizar a velocidade em que recebem, analisam e agem sobre ela —, é de uma pesquisa telefônica realizada pelo Bureau of Labor Statistics [Bureau de Estatísticas de Trabalho] e a informação refere-se a três semanas antes — ou 2 bilhões de milissegundos — do momento em que é divulgada.

Quando empresas gastam milhões de dólares para cortar milissegundos do fluxo de informação, pode parecer um tanto estranho que o governo demore tanto tempo para calcular a taxa de desemprego.

Na verdade, obter esses números cruciais mais cedo foi uma das principais pautas de Alan Krueger ao se tornar membro do Conselho de Economia do Presidente Obama em 2011. Ele não teve sucesso. “Ou o BLS não tem os recursos”, concluiu ele. “Ou eles estão presos na mentalidade do século XX.”

Com o governo claramente não tomando providências para ajustar o passo tão logo, existe uma maneira de pelo menos obter uma mensuração aproximada da taxa de desemprego mais rápido? Na era de alta tecnologia em que vivemos — quando praticamente todos os cliques na internet são registrados em algum lugar —, realmente precisamos esperar semanas para descobrir quantas pessoas estão desempregadas?

Uma potencial solução foi inspirada no trabalho de um ex-engenheiro da Google, Jeremy Ginsberg. Ginsberg percebeu que os dados de saúde, assim como os de desemprego, eram divulgados com demora pelo governo. Os Centros de Prevenção e Controle de Doenças demoram uma semana para divulgar os dados da gripe, apesar do enorme benefício para médicos e hospitais receberem os dados o quanto antes.

Ginsberg suspeitou que pessoas com gripe deveriam fazer buscas relacionadas à gripe. Na essência, reportariam seus sintomas para o Google. Essas buscas poderiam oferecer uma mensuração razoável da taxa atual de gripe. De fato, buscas como “sintoma de gripe” e “dores musculares” se mostraram importantes indicadores da velocidade de disseminação da gripe.*

Enquanto isso, os engenheiros da Google criaram um serviço, o Google Correlate, que dá aos pesquisadores externos os meios para experimentar o mesmo tipo de análise em uma ampla gama de áreas, não apenas na saúde. Os pesquisadores podem pegar qualquer série de dados que estiverem rastreando e verificar quais buscas no Google se correlacionam mais com aquele conjunto de dados.

Por exemplo, usando o Google Correlate, Hal Varian, diretor de economia da Google, e eu fomos capazes de demonstrar quais buscas se relacionam mais intimamente aos preços de moradia. Quando os preços de moradia estão subindo, os norte-americanos tendem a buscar expressões como "hipoteca 80/20", "construtoras" e "taxa de valorização". Quando os preços de moradia estão caindo, os norte-americanos buscam termos como "processo de venda a descoberto", "dívida de hipoteca superior ao imóvel" e "perdão de dívida de hipoteca".

Assim, as buscas no Google funcionam como um teste de contraste para o desemprego da mesma maneira que no caso da gripe ou dos preços de moradia? Podemos dizer, simplesmente através das buscas das pessoas no Google, quantas estão desempregadas, e podemos fazer isso antes que o governo coteje os resultados das pesquisas?

Um dia, inseri as taxas de desemprego dos Estados Unidos entre 2004 e 2011 no Google Correlate.

Dos trilhões de buscas no Google durante este período, o que você acha que se mostrou mais intimamente conectado ao desemprego? Pode pensar que foi "seguro-desemprego" — ou algo do tipo. Esta foi alta, mas não uma das mais altas. "Novos empregos"? Também foi elevada, mas não estava entre as mais frequentes.

O termo mais buscado durante o período que pesquisei — e esses termos tendem a mudar — foi "Slutload". Isso mesmo, a busca mais frequente foi por um site de pornografia. Isso pode parecer estranho a princípio, mas pessoas desempregadas presumidamente têm muito tempo livre. Muitas ficam presas em casa, sozinhas e entendiadas. Outra busca altamente correlacionada — essa no PG realm — é "Spider Solitaire" [paciência]. Novamente, não é surpreendente para um grupo de pessoas que supostamente tem muito tempo disponível.

Não estou alegando, com base nesta única análise, que rastrear buscas por "Slutload" ou "Spider Solitaire" seja a melhor maneira de prever a taxa de desemprego. As distrações específicas que pessoas desempregadas utilizam podem mudar ao longo do tempo (em determinado momento, "Rawtube", um outro site de pornografia, estava ente as correlações mais fortes), e nenhum destes termos em particular sozinhos induz a qualquer coisa que se aproxime de uma pluralidade de desempregados. Mas descobri que normalmente uma mistura de buscas relacionadas a distrações podem rastrear a taxa de desemprego — e seria uma parte do melhor modelo para a prever.

Este exemplo ilustra o primeiro poder do Big Data, a reconsideração do que se qualifica como dado. Frequentemente, o valor do Big Data não é o tamanho, e sim sua capacidade de oferecer novos tipos de informações para estudo — que nunca antes haviam sido coletadas.

Antes do Google havia informações disponíveis sobre certas atividades de lazer — vendas de ingressos de teatro, por exemplo — que poderiam oferecer algumas pistas de quanto tempo livre as pessoas têm. Mas a oportunidade de saber o quanto as pessoas jogam paciência ou assistem a pornografia é nova — e poderosa. Neste exemplo, os dados podem nos ajudar a mensurar mais rapidamente como a economia está — pelo menos até que o governo aprenda a realizar e checar uma pesquisa com maior rapidez.

A vida no campus da Google, em Mountain View, Califórnia, é muito diferente daquela nos escritórios do Goldman Sachs, em Manhattan. Às 9 da manhã os escritórios da Google estão quase vazios. Se houver qualquer funcionário por lá, provavelmente está tomando o café da manhã gratuito — panquecas de banana e mirtilo, ovos mexidos, água aromatizada com pepino. Alguns funcionários podem estar fora da cidade: em uma reunião externa em Boulder ou Las Vegas ou ainda em uma viagem gratuita para esquiar em Lake Tahoe. Por volta do horário do almoço, as quadras de vôlei de praia e os campos de futebol society estão lotadas. O melhor burrito que já comi foi no restaurante mexicano da Google.

Como é possível uma das empresas de alta tecnologia mais competitivas do mundo parecer tão relaxada e generosa? A Google emprega Big Data de uma forma que nenhuma outra empresa jamais fez para construir um fluxo automático de dinheiro. A empresa desempenha um papel crucial neste livro, considerando que os dados de busca do Google são de longe a fonte predominante de Big Data. Mas é importante lembrar que o sucesso do Google é também construído sobre a coleta de um novo tipo de dados.

Se você tem idade suficiente para ter usado a internet no século XX, deve se lembrar dos diversos mecanismos de buscas que existiam na época — MetaCrawler, Lycos, AltaVista, para citar alguns. E deve lembrar que esses mecanismos de busca eram, se tanto, pouco confiáveis. Às vezes, se tivesse sorte, eles conseguiam encontrar o que você desejava. Com frequência, não encontravam. Se digitasse "Bill Clinton" nos mecanismos de busca mais populares do final dos anos 1990, os principais resultados incluiriam sites aleatórios que simplesmente citavam

“Bill Clinton é uma droga” ou um site que contivesse uma piada com Clinton. O que dificilmente pode ser considerada a informação mais importante sobre o então presidente dos Estados Unidos.

Em 1998, surgiu o Google. E seus resultados de busca foram inegavelmente melhores do que todos os de seus outros concorrentes. Se digitasse “Bill Clinton” no Google em 1998, os resultados incluiriam seu site, o endereço de e-mail da Casa Branca e as melhores biografias sobre ele disponíveis na internet. O Google parecia mágico.

O que os fundadores do Google, Sergey Brin e Larry Page fizeram de diferente?

Outros mecanismos de busca localizavam para seus usuários os sites que continham mais ocorrências da frase pela qual buscaram. Se estivesse procurando por informação sobre “Bill Clinton”, aqueles mecanismos de busca encontrariam, pela internet inteira, os sites que continham mais ocorrências para Bill Clinton. Havia muitos motivos para esse sistema de classificação ser imperfeito, e um deles era o fato de ser muito fácil manipulá-lo. Um site de piada com o texto “Bill Clinton Bill Clinton Bill Clinton Bill Clinton Bill Clinton” escondido em algum lugar de sua página seria melhor classificado do que o site oficial da Casa Branca.*

O que Brin e Page fizeram foi encontrar uma forma de registrar um novo tipo de informação muito mais valiosa do que uma mera contagem de palavras. Sites normalmente, ao discutir um assunto, criam links para os sites que entendem ser mais úteis para o entendimento do assunto. Por exemplo, o *New York Times*, se mencionasse Bill Clinton, poderia permitir que os leitores clicassem no nome para ser direcionado para o site oficial da Casa Branca.

Cada site criando esses tipos de links estava, de certa forma, fornecendo uma opinião sobre onde encontrar a melhor informação sobre Bill Clinton. Brin e Page conseguiram reunir todas essas opiniões sobre todos os assuntos. Isso permitiu processar as opiniões do *New York Times*, milhões de Listservs, centenas de bloggers e todo o restante da internet. Se todo um grupo de pessoas pensasse que o link mais importante para “Bill Clinton” fosse o seu site oficial, esse provavelmente seria o site que a maioria das pessoas buscando por “Bill Clinton” gostaria de visualizar.

Esses tipos de links eram dados que outros mecanismos de busca ignoravam, mas que eram incrivelmente preditivos sobre a informação mais útil sobre determinado tópico. A questão aqui é que o Google não dominou as buscas apenas coletando mais dados do que qualquer outro. Ele fez isso encontrando um tipo melhor de dados. Pouco mais de dois anos depois de seu lançamento, o Google, impulsionado por sua análise de links, tornou-se o mecanismo de busca mais popular da internet. Hoje, Brin e Page valem juntos mais de US$60 bilhões.

É assim com o Google e com todo mundo que tenta usar os dados para entender o mundo. A revolução do Big Data tem menos relação com coletar cada vez mais dados e mais com coletar os dados certos.

Mas a internet não é o único lugar em que você pode coletar novos dados e em que os obter traz resultados profundamente discordantes/desafiadores. Este livro é principalmente sobre como os dados na internet podem nos ajudar a entender melhor as pessoas. A próxima seção, no entanto, não tem nada a ver com pessoas. Mas ajuda a ilustrar a principal ideia deste capítulo: o gigantesco valor dos novos dados, não convencionais. E os princípios que isso nos ensina são muito úteis na compreensão da revolução de dados em base digital.

CORPOS COMO DADOS

No verão de 2013, um cavalo alazão de crina preta e de tamanho abaixo da média descansava em uma pequena cocheira ao norte do estado de Nova York. Ele era um dos 152 cavalos de um ano de idade no leilão de agosto da “Fasig-Tipton Select Yearling Sale” em Saratoga Springs, e um dos milhares de cavalos sendo vendidos naquele ano.

Mulheres e homens abastados, quando desembolsam muito dinheiro em um cavalo de corrida, querem a honra de escolher seu nome. Assim, o pequeno alazão ainda não tinha nome, como a maioria dos outros cavalos do leilão, e era chamado apenas pelo número de sua baia, 85.

Não havia muito para parecer que o n° 85 se destacasse neste leilão. Seu pedigree era bom, mas não excelente. Seu pai, *Pioneerof* [sic] *the Nile*, era um excelente cavalo de corrida, mas outros filhos de *Pioneerof the Nile* não alcançaram muito sucesso nas corridas. Havia ainda dúvidas em razão da aparência do n° 85. Ele tinha um arranhão em seu tornozelo, por exemplo, que alguns compradores poderiam achar que se tratava do vestígio de uma lesão.

O atual proprietário do n° 85 era o magnata da cerveja egípcio, Ahmed Zayat, que havia ido ao leilão para vender o cavalo e comprar alguns outros.

Assim como todos os proprietários, Zayat contratou uma equipe de especialistas para o ajudar a escolher que cavalos comprar. Mas seus especialistas eram um tanto diferentes dos tipos contratados por quase todos os proprietários. Os especialistas em cavalos típicos de eventos como esses seriam homens de meia-idade, muitos vindos de Kentucky ou da área rural da Flórida, com baixa instrução, mas com histórico familiar no ramo de cavalos. Os especialistas de Zayat, porém, eram de uma pequena empresa chamada EQB. O chefe da EQB não era

um típico especialista em cavalos. Ele era Jeff Seder, um homem excêntrico nascido na Filadélfia com uma pilha de diplomas de Harvard.

Zayat já havia trabalhado com a EQB, então o processo já lhe era familiar. Depois de alguns dias avaliando os cavalos, a equipe de Seder entregaria para Zayat mais ou menos cinco cavalos recomendados para compra para substituir o n° 85.

Desta vez, porém, foi diferente. A equipe de Seder reportou a Zayat que não poderia atender a seu pedido. Eles não poderiam recomendar que ele comprasse qualquer dos outros 151 cavalos oferecidos no leilão. Além disso, fizeram um pedido súbito e quase desesperado. Zayat em hipótese alguma deveria vender o n° 85. O cavalo, declarou a EQB, não apenas era o melhor do leilão, como era o melhor cavalo do ano e, possivelmente, da década. “Venda sua casa”, suplicou a equipe. “Mas não venda esse cavalo.”

No dia seguinte, com pouco alarde, o Cavalo n° 85 foi comprado por US$300 mil por um homem chamado Incardo Bloodstock. Bloodstock, mais tarde revelaram, foi o pseudônimo usado por Ahmed Zayat. Em resposta às suplicas de Seder, Zayat comprou o próprio cavalo, uma ação raramente presenciada. (As regras do leilão impediam que Zayat simplesmente retirasse o cavalo, por isso ele precisou fazer a transação usando um pseudônimo.) Sessenta e dois cavalos foram vendidos no leilão por um preço mais alto do que o n° 85, dois deles atingiram mais de US$1 milhão cada.

Três meses depois, Zayat finalmente escolheu um nome para o n° 85: *American Pharoah*. E dezoito meses depois, em uma noite agradável de sábado na cidade de Nova York, *American Pharoah* tornou-se o primeiro cavalo em mais de três décadas a vencer a Tríplice Coroa.

O que Jeff Seder sabia sobre o Cavalo n° 85 que aparentemente ninguém mais sabia? Como esse egresso de Harvard ficou tão bom em avaliar cavalos?

Conheci Seder, que na época tinha 64 anos, em uma escaldante tarde de junho em Ocala, Flórida, mais de 5 anos após a vitória da Tríplice Coroa por *American Pharoah*. O evento era uma exposição de uma semana para cavalos de 2 anos, culminando com um leilão, não muito diferente do evento de 2013 em que Zayat comprara seu cavalo de volta.

Seder tinha uma voz estrondosa, ao estilo Mel Brooks, vastos cabelos e uma cadência peculiar ao andar. Ele vestia suspensórios, calças cáqui, uma camisa preta com o logo de sua empresa e um aparelho auditivo.

Ao longo de três dias, ele me contou sua história de vida — e como se tornou um excelente avaliador de cavalos. Não foi bem um caminho natural. Depois de se graduar com honras em Harvard, Seder partiu para a graduação, também em Harvard, em direito e em negócios. Aos 26 anos, trabalhava como analista no Citigroup, em Nova York, mas estava infeliz e esgotado. Um dia, sentado no átrio do novo escritório da empresa na Lexington Avenue, viu-se analisando um enorme mural de um terreno baldio. A pintura lhe fez recordar de seu amor pela vida no campo e pelos cavalos. Foi para casa e se olhou no espelho vestindo um terno de três peças. Naquele momento ele soube que não fora talhado para trabalhar em um banco e viver em Nova York. Na manhã seguinte, pediu demissão.

Seder se mudou para a área rural da Pensilvânia, e passou por uma variedade de empregos na área têxtil e da medicina esportiva antes de se dedicar em tempo integral à sua paixão: prever o sucesso de cavalos de corrida. Os números nas corridas de cavalos são brutais/aproximados. Dentre os mil cavalos de dois anos apresentados no leilão de Ocala, um dos mais prestigiados do país, talvez cinco acabem vencendo uma corrida com um prêmio significativo. O que acontecerá com os outros 995 cavalos? Aproximadamente um terço será lento demais. Outro terço sofrerá lesões — a maioria porque as patas não conseguem suportar a enorme pressão de galopar a toda velocidade. (A cada ano, centenas de cavalos morrem em pistas de corrida nos Estados Unidos, grande parte por fraturas nas pernas.) E o terço restante sofrerá o que podemos chamar de Síndrome de Bartleby. Bartleby, um escrivão do extraordinário conto de Herman Melville, para de trabalhar e responde a todas as perguntas de seu chefe com: “Prefiro não fazer.” Muitos cavalos, no início de sua carreira de corredor, aparentemente se dão conta de que não precisam correr se não quiserem. Começam a corrida correndo muito, mas, em determinado ponto, simplesmente diminuem a velocidade ou param completamente de correr. Para que correr em uma pista oval o mais rápido que puder, especialmente se seus cascos e jarretes doem? “Prefiro não”, decidem. (Tenho um ponto fraco para Bartlebys, cavalos ou humanos.)

Com as probabilidades contra eles, como os proprietários podem identificar um cavalo lucrativo? Historicamente, as pessoas acreditavam que a melhor maneira de prever se um cavalo teria sucesso era analisando seu pedigree. Ser um especialista em cavalos significava ser capaz de recitar tudo que alguém poderia querer saber sobre pai, mãe, avó, avô, irmãos e irmãs do cavalo. Agentes costumam declarar, por exemplo, que um cavalo grande “atingiu seu tamanho legítimo” quando a linhagem de sua mãe tem muitos cavalos grandes.

Entretanto, existe um problema. Embora o pedigree seja importante, só consegue explicar uma pequena parte do sucesso de cavalos de corrida. Considere os tempos de pista de irmãos de pai e mãe de todos os cavalos que

venceram o prêmio anual de *Cavalo do Ano*, o mais prestigiado no mundo das corridas. Esses cavalos têm os melhores pedigrees possíveis — histórico familiar idêntico ao dos maiores cavalos da história. Ainda assim, mais de três quartos não vencem uma grande corrida. A forma tradicional de prever o sucesso de um cavalo, pelo que os dados nos dizem, deixam muita margem para aprimoramento.

Na verdade, não é surpreendente que um pedigree não seja tão preditivo. Pense em pessoas. Imagine um dono de um time da NBA que comprasse seu futuro time aos dez anos de idade com base em seu pedigree. Ele contraria um agente para avaliar Earvin Johnson III, filho de "Magic" Johnson. "Ele tem uma altura boa até agora", diria o agente. "Tem o tamanho legítimo, da linhagem de Johnson. Ele deve ter uma ótima visão, abnegação, altura e velocidade. Parece sociável, com ótima personalidade. Andar confiante. Bem-apessoado. Ele é uma ótima aposta." Infelizmente, catorze anos depois, o dono do time teria um blogueiro de moda do canal E! de 1,88m (baixo para um jogador de basquete profissional). Earvin Johnson III poderia ser de grande ajuda para desenhar os uniformes do time, mas provavelmente, dispensável em uma quadra.

Junto com o blogueiro de moda, o dono do time da NBA que escolhera a equipe da mesma forma que muitos donos escolhem seus cavalos provavelmente teria recrutado Jeffrey e Marcus Jordan, filhos de Michael Jordan, ambos que se provaram jogadores universitários medíocres. Boa sorte contra o Cleveland Cavaliers, que conta com a estrela LeBron James, cuja mãe tem 1,65m de altura. Ou imagine um país que eleja seus líderes com base em seus pedigrees. Teríamos presidentes como George W. Bush. (Desculpe, não resisti.)

Agentes de cavalos usam outras informações, além do pedigree. Por exemplo, analisam a andadura de potros de dois anos e examinam visualmente o cavalo. Em Ocala, passei horas conversando com diversos agentes, o que foi suficiente para verificar que há pouco consenso sobre o que de fato procuram em um cavalo.

Acrescente a essas gigantescas contradições e incertezas o fato de que alguns compradores de cavalos parecem ter recursos infinitos, e terá um mercado com consideráveis ineficiências. Dez anos atrás, o Cavalo n° 153, aos dois anos, corria mais rápido do que qualquer outro, era exuberante para a maioria dos agentes e tinha um fantástico pedigree — descendente de *Northern Dancer* e *Secretariat*, dois dos maiores cavalos de corrida de todos os tempos. Tanto um bilionário irlandês quanto um xeique de Dubai queriam comprá-lo. Eles entraram em uma batalha de lances no leilão que rapidamente se transformou em um concurso de egos. Enquanto centenas de homens e mulheres observavam estupefatos, os lances continuaram subindo, até que o potro de dois anos foi finalmente vendido por US$16 milhões, de longe o maior preço pago por um cavalo. O Cavalo n° 153, que recebeu o nome de *The Green Monkey*, participou de três corridas, ganhou apenas US$10 mil e foi aposentado.

Seder nunca teve qualquer interesse pelos métodos tradicionais de avaliação de cavalos. Ele só se interessava por dados. Planejou mensurar diversos atributos de cavalos de corrida e ver qual deles se correlacionava com seu desempenho. É importante observar que Seder trabalhou nesse plano cinco anos antes do surgimento da internet. Mas sua estratégia foi embasada pela ciência de dados. E as lições colhidas de sua história são aplicáveis a qualquer um que use Big Data.

Durante anos, a busca de Seder lhe rendeu apenas frustração. Ele mediu o tamanho das narinas dos cavalos, criando o primeiro e maior conjunto de dados do mundo sobre o tamanho de narinas de cavalos e eventuais lucros obtidos. Ele descobriu que o tamanho da narina não era um preditor do sucesso do cavalo. Então, submeteu os cavalos a eletrocardiogramas e coletou os membros de cavalos mortos para medir o volume dos músculos de contração rápida. Certa vez, analisou o tamanho dos excrementos dos cavalos, investigando a teoria de que perder muito peso antes de uma corrida pudesse tornar o cavalo mais lento. Nenhum desses fatores se correlacionou ao sucesso do animal.

Então, vinte anos atrás, ele teve sua primeira grande descoberta. Seder decidiu medir o tamanho dos órgãos internos dos cavalos. Como isso era impossível com a tecnologia existente na época, construiu o próprio ultrassom portátil. Os resultados foram impressionantes. Ele descobriu que o tamanho do coração, e especialmente do ventrículo esquerdo, era um enorme preditor do sucesso de um cavalo, a variável única mais importante. Outro órgão que fazia diferença era o baço: cavalos com baço pequeno rendiam praticamente nada.

Seder teve alguns outros êxitos. Digitalizou milhares de vídeos de cavalos galopando e descobriu que certos gestos de galope se correlacionavam com o sucesso nas pistas. Também descobriu que alguns potros de dois anos assobiam ao respirar depois de correr duzentos metros. Esses cavalos às vezes eram vendidos por até um milhão de dólares, mas os dados de Seder lhe mostraram que os "assobiadores" quase nunca tinham sucesso. Assim, ele posicionou um assistente perto da linha de chegada para identificar os "assobiadores".

Dentre os cerca de mil cavalos no leilão de Ocala, aproximadamente dez passariam em todos os testes de Seder. Ele ignora completamente o pedigree, exceto como fator de influência no preço por que o cavalo será vendido. "O pedigree nos mostra que um cavalo pode ter uma chance bem pequena de ser excelente", diz. "Mas se vejo que o cavalo é bom, que importa como chegou lá?"

Uma noite, Seder me convidou para seu quarto no Hotel Hilton, em Ocala. Lá, me contou sobre sua infância, família e carreira. Ele me mostrou fotos de sua esposa, sua filha e seu filho. Contou que era um entre apenas três alunos judeus em sua escola de ensino médio na Filadélfia, e que quando ingressou no colégio tinha 1,47m. (Na faculdade chegou a 1,75m.) Falou de seu cavalo favorito: Pinky Pizwaanski. Seder comprou e batizou o cavalo em homenagem a um jóquei homossexual. Sentia que Pinky, o cavalo, sempre se esforçava bastante mesmo não sendo muito bem-sucedido.

Finalmente, me mostrou um arquivo que continha todos os dados que registrara sobre o n° 85, o arquivo responsável pela maior predição de sua carreira. Será que revelaria seu segredo? Talvez, mas disse que não se importava. Mais importante do que proteger seus segredos era provar que estava correto, mostrando ao mundo que os vinte anos dissecando membros, coletando fezes e construindo ultrassons improvisados tinham valido a pena.

Veja alguns dados sobre o Cavalo n° 85:

PERCENTIS DO N° 85 (POSTERIORMENTE *AMERICAN PHAROAH*) COM UM ANO

	PERCEN
Altura	56
Peso	61
Pedigree	70
Ventrículo esquerdo	*99,61*

Lá estava, totalmente claro, o motivo de Seder e sua equipe se tornarem tão obcecados pelo n° 85. Seu ventrículo esquerdo estava no percentil de 99,61!

Não era apenas isso, mas todos os outros órgãos importantes, incluindo o restante do coração e o baço, eram excepcionalmente grandes também. Em termos gerais, quando se trata de corridas, Seder descobriu que quanto maior o ventrículo esquerdo, melhor. Mas um ventrículo esquerdo desse tamanho pode ser indicação de doenças se os outros órgãos forem pequenos. No *American Pharoah*, todos os órgãos-chave eram maiores que a média, e o ventrículo esquerdo era enorme. Os dados gritavam que o n° 85 era 1 em 100 mil ou até 1 em 1 milhão.

O que os cientistas de dados podem aprender com o projeto de Seder?

Primeiro, e talvez o mais importante, se pretende utilizar novos dados para revolucionar um campo, é melhor tentar em um onde os métodos antigos são ruins. Os agentes de cavalos obcecados por pedigree e superados por Seder deixaram muito espaço para aprimoramento. Assim como os mecanismos de buscas obcecados por contagem de palavras superados pelo Google.

Uma fraqueza da tentativa do Google de prever a gripe usando dados de busca é que já é possível prevê-la muito bem com os dados da semana anterior e um simples ajuste sazonal. Ainda há debates sobre o quanto esses dados são capazes de acrescentar a esse modelo simples e poderoso. Em minha opinião, as buscas no Google são mais promissoras medindo as condições de saúde em que os dados existentes são fracos, e, portanto, algo como o Google padrão pode se mostrar mais valioso em um período mais longo do que o Google Flu.

A segunda lição é que, ao tentar fazer predições, você não precisa se preocupar demais no porquê de seu modelo funcionar. Seder não poderia explicar inteiramente por que o ventrículo esquerdo é tão importante na predição do sucesso de um cavalo. Nem poderia calcular com precisão o valor do baço. Talvez um dia cardiologistas e hematologistas veterinários resolverão este mistério. Mas por ora isso não importa. Seder está no campo da predição, não no da explicação. E, na predição, você só precisa saber que algo funciona, não o porquê.

Por exemplo, o Walmart utiliza dados das vendas em todas as lojas para saber quais produtos estocar. Antes do Furacão Frances, uma terrível tempestade que atingiu o Sudeste dos Estados Unidos em 2004, o Walmart suspeitou — corretamente — que os hábitos de compra das pessoas mudavam quando a cidade estava prestes a ser atingida por uma tempestade. Eles analisaram os dados de vendas de furacões anteriores para saber o que as pessoas comprariam. Uma resposta predominante? Pop-Tarts de morango. Este produto vende sete vezes mais rápido do que o normal em dias que antecedem furacões.

Com base em suas análises, o Walmart providenciou o envio de caminhões carregados de Pop-Tarts de morango para as lojas no trajeto do furacão. E de fato essas Pop-Tarts venderam bem.

Por que Pop-Tarts? Provavelmente porque não precisam de refrigeração ou cozimento. Por que morango? Não se sabe. Mas quando furacões chegam, as pessoas aparentemente recorrem a Pop-Tarts de morango. Então, nos dias que antecedem um furacão, o Walmart agora regularmente estoca caixas e mais caixas de Pop-Tarts de morango. A razão para essa correlação não importa. Mas sim a relação em si. Talvez algum dia cientistas de alimentos desvendem a associação entre furacões e tortas recheadas com geleia de morango. Mas enquanto espera por uma explicação, o Walmart ainda precisa estocar e lotar suas prateleiras com Pop-Tarts de morango quando há furacões a caminho, e guarda o Rice Krispies para dias mais ensolarados.

Esta lição fica nítida também na história de Orley Ashenfelter. O que Seder representa para os cavalos, Ashenfelter, economista de Princeton, representa para os vinhos.

Há pouco mais de uma década, Ashenfelter estava frustrado. Ele comprava muitos vinhos tintos da região de Bordeaux, na França. Às vezes, o vinho era delicioso, valia o alto preço pago. Muitas vezes, porém, era uma decepção.

Por que, Ashenfelter ponderava, estava pagando o mesmo preço por vinhos com avaliações tão diferentes?

Um dia, Ashenfelter recebeu uma dica de um amigo jornalista e conhecedor de vinhos. Havia de fato uma maneira de descobrir se um vinho seria bom. O segredo, disse o amigo de Ashenfelter, era o clima durante a estação de crescimento da uva.

O interesse de Ashenfelter se intensificou. Foi então quando iniciou sua jornada para descobrir se isso era verdade e se conseguiria comprar de modo consistente vinhos melhores. Ele baixou trinta anos de dados climáticos da região de Bordeaux. Também coletou os preços em leilões de vinho. Esses leilões, que ocorrem muitos anos após o vinho ter sido originalmente vendido, poderiam ser um bom indicador se o vinho se mostrou bom.

O resultado foi fantástico. Um grande porcentual da qualidade de um vinho era explicado simplesmente pelo clima durante a estação de crescimento.

Na verdade, a qualidade de um vinho se resumia em uma fórmula simples, que poderíamos chamar de Primeira Lei da Viticultura:

> Preço = 12,145 + 0,00117 índice pluviométrico no inverno + 0,0614 temperatura média da estação de crescimento – 0,00386 índice pluviométrico na colheita.

Então por que a qualidade do vinho em Bordeaux funciona assim? O que explica a Primeira Lei da Viticultura? Existe alguma explicação para a fórmula do vinho de Ashenfelter — calor e irrigação no início são essenciais para um amadurecimento adequado.

Mas os detalhes precisos desta fórmula preditiva vão muito além de qualquer teoria, e ela provavelmente nunca será plenamente entendida pelos especialistas no campo.

Por que um centímetro de chuva no inverno acrescenta, em média, exatamente US$0,1 centavo ao preço de uma garrafa madura de vinho tinto? Por que não US$0,2 centavos? Por que não US$0,05? Ninguém pode responder essas perguntas. Mas se houver 1.000 centímetros de chuva extra no inverno, você deve estar disposto a pagar US$1 a mais pela garrafa de vinho.

De fato, Ashenfelter, apesar de não saber exatamente por que sua regressão funcionava, a usou para comprar vinhos. De acordo com ele: “Funcionou muito bem.” A qualidade dos vinhos que bebeu foi perceptivelmente melhor.

Se seu objetivo é prever o futuro — qual vinho será saboroso, quais produtos irá vender, quais cavalos serão mais velozes —, você não precisa se preocupar muito sobre o porquê exatamente seu modelo funciona. Basta obter os números certos. Esta é a segunda lição da história de Jeff Seder com os cavalos.

A última lição a ser aprendida da tentativa bem-sucedida de Seder em prever o potencial vencedor da Tríplice Coroa é que é preciso ser flexível na determinação do que se classifica como dados. Não é que os antigos agentes de cavalos desconhecem os dados antes de Seder. Eles esmiuçavam os tempos de corrida e os mapas de pedigrees. A genialidade de Seder foi buscar os dados onde os outros sequer olharam, considerar suas fontes não tradicionais. Para o cientista de dados, uma perspectiva nova e original pode valer a pena.

PALAVRAS COMO DADOS

Certo dia em 2004, dois jovens economistas especializados em mídia, na época estudantes de doutorado em Harvard, liam sobre uma recente decisão do tribunal de Massachusetts legalizando o casamento gay.

Os economistas, Matt Gentzkow e Jesse Shapiro, notaram algo interessante: dois jornais empregavam linguagens visivelmente diferentes para relatar a mesma história. O *Washington Times*, que tem a reputação de ser

conservador, escreveu a seguinte manchete: “Homossexuais ‘Se Casam’ em Massachusetts.” O *Washington Post*, que tem a reputação de ser liberal, relatou uma vitória para os “casais do mesmo sexo”.

Não é de surpreender que diferentes organizações de notícias tendam a direções distintas, que jornais cubram a mesma matéria de um ângulo diferente. Durante anos, na verdade, Gentzkow e Shapiro ponderaram se usariam seu conhecimento em economia para entender a tendência da mídia. Por que algumas empresas de notícias parecem assumir uma perspectiva mais liberal e outras, uma mais conservadora?

Mas Gentzkow e Shapiro não tinham ideia de como abordar esta questão; não conseguiam descobrir como mediriam de forma sistemática e objetiva a subjetividade da mídia.

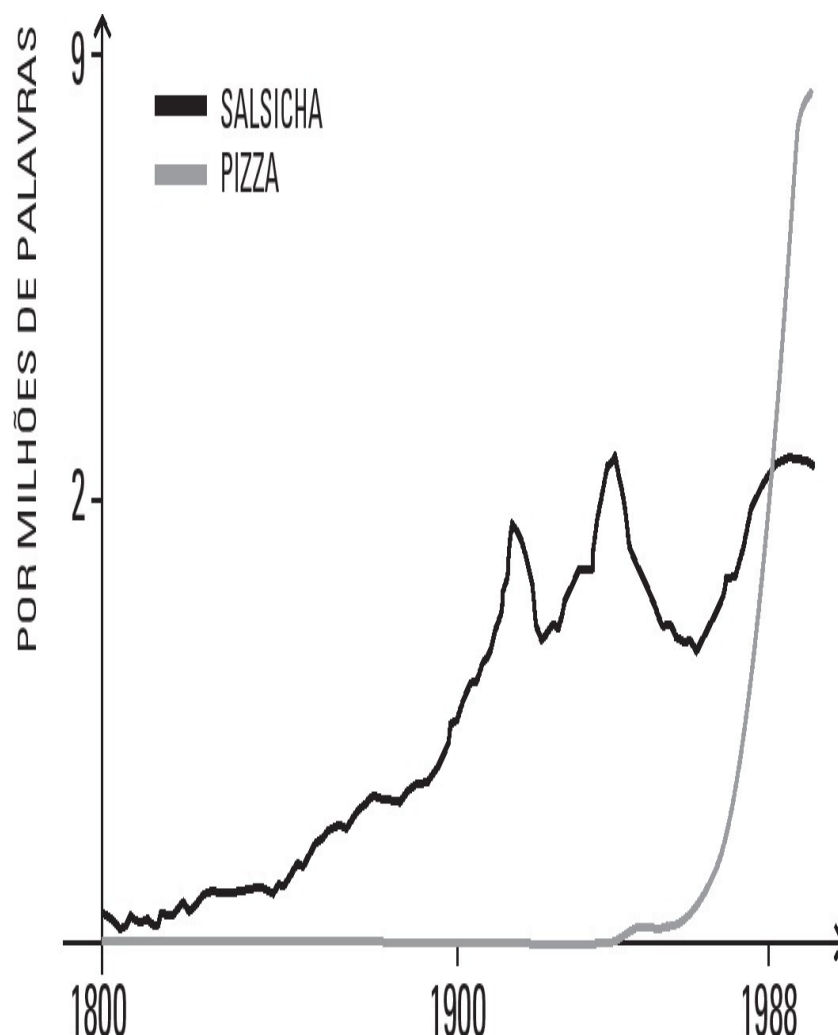
O que Gentzkow e Shapiro acharam interessante sobre a história do casamento gay não foi o fato de os jornais divergirem em suas coberturas, e sim *como* divergiam — tudo se resumia a uma peculiar mudança na escolha de palavras. Em 2004, o termo “homossexuais” usado pelo *Washington Times*, era uma forma antiquada e pejorativa de descrever os gays, enquanto que a expressão “casais do mesmo sexo”, como usado pelo *Washington Post*, enfatizava que relacionamentos gays eram apenas outra forma de amor.

Os acadêmicos se perguntaram se a linguagem poderia ser a chave para entender a tendência. Será que liberais e conservadores sempre usam expressões diferentes? As palavras utilizadas por um jornal em suas matérias poderiam ser transformadas em dados? O que isso revelaria sobre a imprensa norte-americana? Conseguiríamos descobrir se a imprensa era liberal ou conservadora? E será que encontraríamos uma explicação? Em 2004, estas não eram perguntas inúteis. As bilhões de palavras nos jornais norte-americanos não estavam mais aprisionadas em impressos ou microfilmes. Certos sites agora registram cada palavra empregada em cada matéria de qualquer jornal dos Estados Unidos. Gentzkow e Shapiro poderiam, então, analisar esses sites e rapidamente testar a capacidade da linguagem em mensurar a inclinação de um jornal. E, ao fazer isso, conseguiriam aprimorar nossa compreensão de como a mídia de notícias funciona.

Mas, antes de descrever o que eles descobriram, vamos deixar de lado por um instante a história de Gentzkow e Shapiro e sua tentativa de quantificar a linguagem em jornais e discutir como acadêmicos, em uma ampla gama de áreas de conhecimento, utilizaram este tipo de novos dados — palavras — para melhor entender a natureza humana. A linguagem, é claro, sempre foi um assunto de interesse dos cientistas sociais. Entretanto, estudá-la geralmente exigia a leitura minuciosa de textos, e transformar gigantescas pilhas de texto em dados não era viável. Agora, com os computadores e a digitalização, listar palavras contidas em enormes conjuntos de documentos é fácil. A linguagem, assim, se tornou passível de análise de Big Data. Os links utilizados pelo Google eram compostos de palavras. Assim como as buscas no Google estudadas por mim e frequentemente retratadas neste livro. Mas a linguagem é tão importante para a revolução do Big Data que merece uma seção só para ela. Na verdade, ela vem sendo usada tanto que há um campo inteiro dedicado a ela: “texto como dados”.

Um importante desenvolvimento neste campo é o Google Ngrams. Alguns anos atrás, dois jovens biólogos, Erez Aiden e Jean-Baptiste Michel, pediram que seus assistentes de pesquisa contassem palavras, uma por uma, em textos velhos e empoeirados para tentar descobrir novas perspectivas sobre como determinados usos de palavras se disseminam. Certo dia, Aiden e Michel ficaram sabendo de um novo projeto da Google para digitalizar uma grande parte dos livros do mundo. Quase imediatamente, os biólogos perceberam que essa seria uma maneira muito mais fácil de compreender a história da linguagem.

“Percebemos que nossos métodos estavam completamente obsoletos”, contou Aiden à revista *Discover*. “Ficou claro que não dava para competir com essa avalanche de digitalização.” Assim, eles decidiram colaborar com a empresa de buscas. Com a ajuda dos engenheiros da Google, criaram um serviço que faz buscas por uma palavra ou frase específica em milhões de livros digitalizados. Isso então diz aos pesquisadores com que frequência aquela palavra ou frase apareceu a cada ano, de 1800 a 2010.

O que podemos aprender com a frequência com que palavras ou frases apareciam nos livros em diferentes anos? Primeiro, descobrimos sobre o lento crescimento na popularidade da salsicha e o relativamente recente e rápido crescimento na popularidade da pizza.

Mas há lições mais profundas que essa. Por exemplo, o Google Ngrams pode nos ensinar como a identidade nacional é criada. Um exemplo fascinante é apresentado no livro de Aiden e Michel, *Uncharted* ["Inexplorado", em tradução livre].

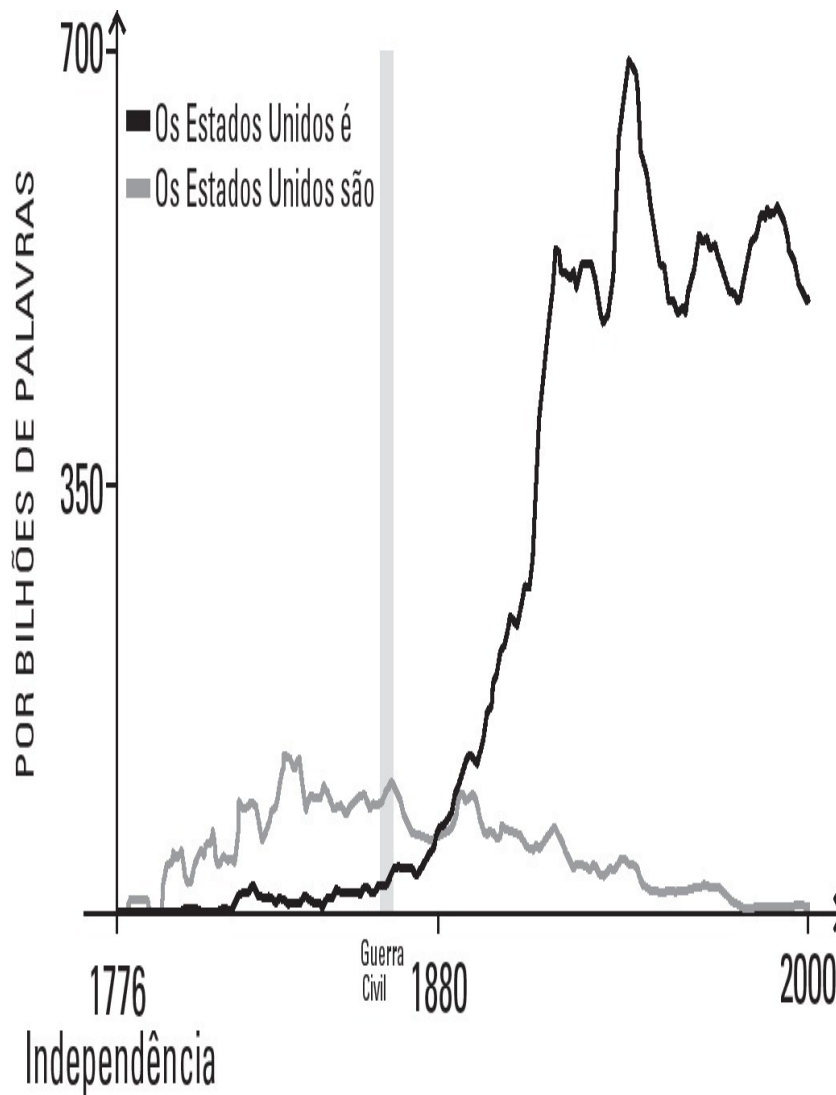
Primeiro, uma rápida pergunta. Você acha que os Estados Unidos atualmente são um país unido ou dividido? Se você é como a maioria das pessoas, dirá que hoje o Estados Unidos é dividido em razão da alta polarização política. Pode até mesmo dizer que o país está mais dividido do que nunca. Os Estados Unidos, afinal, são separados por cores: estados vermelhos são Republicanos; estados azuis, Democratas. Mas em *Uncharted*, Aiden e Michel observam um fascinante ponto de dados que revela o quanto os Estados Unidos já foram mais divididos. O ponto de dados é a linguagem que as pessoas usam para falar sobre o país.

Preste atenção às palavras que usei no parágrafo anterior ao falar sobre o quanto o país está dividido. Escrevi: "O Estados Unidos é dividido." Usei o verbo no singular, me referindo a Estados Unidos como substantivo singular, que é a gramática correta e o uso padrão em inglês.

Entretanto, os norte-americanos nem sempre falaram dessa maneira. Antigamente, usava-se a forma plural. Por exemplo, John Adams, em seu discurso sobre o Estado da União de 1799, se referiu a: "*os* Estados Unidos em seus tratados com a Coroa Britânica." Se estivéssemos conversando em 1800, em inglês, usaríamos a forma plural: "Os Estados Unidos *são* divididos." Esta pequena diferença nas convenções de uso do idioma fascina historiadores há anos, pois sugere que houve um momento em que os Estados Unidos deixaram de pensar em si mesmos como um conjunto de estados e passaram a se considerar uma nação.

Mas quando isto aconteceu? Historiadores, como nos informa o livro *Uncharted*, nunca puderam ter certeza, pois não havia uma maneira sistemática de se verificar. Mas muitos sempre suspeitaram que a causa teria sido a Guerra Civil. Na verdade, James McPherson, ex-presidente da American Historical Association e vencedor do

prêmio Pulitzer, declarou com todas as letras: “A guerra marcou a transição dos Estados Unidos para um substantivo singular.”

Ocorre que McPherson estava errado. O Google Ngrams forneceu a Aiden e Michel uma maneira sistemática de testar a teoria. Eles poderiam analisar com que frequência os livros norte-americanos usaram a frase “Os Estados Unidos são…” versus “O Estados Unidos é…” em cada ano da história do país. A transformação foi mais gradual e não ganhou velocidade até muito tempo depois do fim da Guerra Civil.

Quinze anos depois da Guerra Civil, ainda havia mais usos de “Os Estados Unidos são…” do que “O Estados Unidos é…”, mostrando que o país ainda era linguisticamente dividido. As vitórias militares foram mais rápidas do que a mudança da mentalidade.

É assim que uma nação se une. Mas como um homem e uma mulher se unem? As palavras podem nos ajudar aqui também.

Por exemplo, podemos prever se um homem e uma mulher terão um segundo encontro com base em como falam no primeiro.

Isso foi demonstrado por uma equipe interdisciplinar das universidades de Stanford e Northwestern: Daniel McFarland, Dan Jurafsky e Craig Rawlings. Eles estudaram centenas de encontros rápidos entre heterossexuais e tentaram determinar o que poderia prever se o casal sentiu uma conexão e gostaria de um segundo encontro.

Eles primeiro usaram os dados tradicionais. Perguntaram aos participantes sua altura, peso e hobbies, então, testaram como estes fatores se correlacionavam com alguém que relatara sentir uma fagulha de interesse romântico. Mulheres, na média, preferem homens mais altos e que compartilham os mesmos hobbies; homens, na média, preferem mulheres mais magras e que compartilham os mesmos hobbies. Nada de novo aqui.

Mas os cientistas coletaram um novo tipo de dados. Instruíram os participantes a levar gravadores. As gravações dos encontros foram então digitalizadas. Assim, os cientistas foram capazes de codificar palavras usadas, risadas e tom de voz. Poderiam testar como os homens e as mulheres sinalizavam que estavam interessados e como os parceiros conquistaram este interesse.

O que os dados linguísticos nos disseram? Primeiro, como um homem ou mulher comunica seu interesse. Uma das maneiras de um homem sinalizar que sente atração é óbvia: ele ri das piadas da mulher. Outra é menos óbvia: ao falar, restringe a amplitude do tom de voz. Há pesquisas que indicam que uma voz uniforme é frequentemente vista pelas mulheres como masculina, o que sugere que os homens, talvez de modo subconsciente, exageram sua masculinidade quando gostam de uma mulher.

Os cientistas descobriram que uma mulher sinaliza seu interesse variando o tom de voz, usando um tom mais suave e falando trechos mais curtos. Há também pistas importantes sobre o interesse de uma mulher com base em determinadas palavras que usa. É improvável que uma mulher esteja interessada quando usa palavras e frases como "provavelmente" ou "eu acho".

Camaradas, se uma mulher usar estas palavras — se gosta "mais ou menos" da bebida, "acha" que está com frio ou "pode ser" que queira outro aperitivo — pode apostar que ela "provavelmente" não está nem "mais ou menos" a fim de você.

Uma mulher provavelmente *está* interessada quando fala de si mesma. Ocorre que, para um homem a fim em busca de um sinal, a palavra mais bonita que pode ouvir de uma mulher é: "eu". Isto é um sinal de que ela se sente confortável. Ela provavelmente também está interessada se usar marcadores discursivos, como "sabe" e "quero dizer". Por quê? Os cientistas observaram que estas frases chamam a atenção do interlocutor. Elas são amigáveis e calorosas e sugerem uma pessoa buscando se conectar, sabe o que quero dizer?

Agora, como homens e mulheres podem se comunicar para fazer o parceiro se interessar por eles? Os dados nos mostram que existem diversas maneiras para um homem aumentar suas chances de que a mulher goste dele. Mulheres gostam de homens que seguem suas indicações. Talvez não seja uma surpresa que uma mulher tem mais probabilidade de relatar uma conexão se um homem rir de suas piadas e mantenha a conversa em assuntos que ela introduz em vez de constantemente mudar de assunto para outros sobre o qual queira falar.* Mulheres também gostam de homens que expressam apoio e simpatia. Se um homem diz: "Isso é sensacional!" ou "Isso é muito legal", uma mulher é significativamente mais propensa a relatar haver uma conexão. O mesmo acontece se ele usar frases como: "Nossa, que difícil" ou "Você deve estar triste".

Para as mulheres tenho uma má notícia: os dados parecem confirmar uma verdade detestável sobre os homens. A conversa desempenha apenas um pequeno papel em como reagem às mulheres. A aparência física supera todo o resto como preditor do relato de um homem sobre a conexão no contato. Dito isto, há uma palavra que uma mulher pode usar para pelo menos melhorar um pouquinho suas chances de que o homem goste dela, já falei qual: "eu". Homens são mais propensos a reportar um "clique" com uma mulher que fale sobre si mesma. E como previamente mencionado, uma mulher também tem maior probabilidade de relatar uma conexão quando fala sobre si. Portanto, é um ótimo sinal, em um primeiro encontro, se houver substanciais conversas sobre a mulher. Ela sinaliza seu agrado e provavelmente gosta que o homem não esteja cortando a conversa. E o homem gosta que a mulher fale mais de si. Um segundo encontro é provável.

Finalmente, há um claro indicador de problema na transcrição dos encontros: um ponto de interrogação. Se há muitas perguntas durante o encontro, é menos provável que tanto o homem quanto a mulher relatem ter havido uma conexão. Isto parece contraintuitivo, você pode pensar que perguntas são um sinal de interesse. Mas nem tanto no primeiro encontro, a maioria das perguntas neste momento são sinais de tédio. "Quais são seus hobbies?", "Quantos irmãos e irmãs você tem?" Estes são os tipos de coisas que pessoas dizem quando a conversa empaca. Um ótimo primeiro encontro inclui uma única pergunta no final: "Quer sair comigo de novo?" Se essa for a única pergunta do encontro, a resposta provavelmente será "sim".

E homens e mulheres não apenas falam de modo diferente quando estão tentando atrair o sexo oposto. Eles falam de modo diferente em geral.

Uma equipe de psicólogos analisou as palavras usadas por centenas de milhares de posts no Facebook. Mediram com que frequência cada palavra é usada por homens e mulheres. Eles então declararam quais são as

palavras mais masculinas e femininas em inglês.

Muitas dessas preferências por determinadas palavras, infelizmente, eram óbvias. Por exemplo, mulheres falam sobre “compras” e “meu cabelo” com muito mais frequência do que homens. Eles falam mais sobre “futebol” e “Xbox” do que as mulheres. Você provavelmente não precisaria de uma equipe de psicólogos analisando Big Data para lhe dizer isso.

Algumas das descobertas, porém, foram mais interessantes. Mulheres usam muito mais a palavra “amanhã” do que homens, talvez porque homens não sejam muito bons em pensar adiante. Acrescentar a letra “a” na palavra “tão” é uma das características linguísticas femininas mais fortes. Entre as formas usadas de modo mais desproporcional pelas mulheres estão “tãão”, “tããão”, “tãããããão”, “tãããããão” e “tããããããão”.

Talvez tenha sido minha exposição na infância a mulheres que não tinham qualquer pudor em soltar o verbo, mas sempre pensei que falar palavrão fosse uma característica de ambos os sexos em igualdade de condições. Não é verdade. Entre as palavras usadas com muito mais frequência por homens do que por mulheres estão “merda”, “droga”, “foda-se”, “porra”, “foda” e “fudido”.

As nuvens de palavras a seguir mostram as palavras usadas mais por homens e as por mulheres. Quanto maior a palavra aparece, mais é característica do gênero.

Homens

Mulheres

O que gosto neste estudo é que os novos dados nos informam sobre os padrões que há muito tempo existiam, mas de que não estávamos necessariamente cientes. Homens e mulheres sempre falaram de formas específicas. Mas, há milênios, estes dados desapareceriam assim que as ondas sonoras se dissipassem no espaço. Agora, estes dados são preservados em computadores e podem ser analisados por eles.

Ou talvez o que eu deveria ter dito, considerando meu gênero: “As porras das palavras desapareciam. Agora, podemos dar uma pausa no futebol e no Xbox e aprender essa merda. Isso se alguém se importar com essa porra.”

Não são apenas homens e mulheres que falam de forma diferente. As pessoas usam palavras específicas de acordo com a faixa etária. Isto poderia nos dar algumas pistas sobre o papel do envelhecimento. Veja a seguir, a partir do mesmo estudo, as palavras usadas de modo mais desproporcional por pessoas de diferentes idades no Facebook.

BEBER. TRABALHAR. REZAR.

19 a 22 anos

sair
tipo empresta
curtir não entrar
nsei
algumacoisa pular apostado
correndo emprestar sussa
quer demais
primavera online
horário registrado
crédito aula outono
registro oficialmente
faculdade
indo semestre
aulas entrei
verão
notas final
filosofia prova aula
estudando
ciências
história prof
palestra professor
teste psicologia
psi semestre
<3 bosta
merda vá_para_casa
mulherada
liga_aí colega saudade 21 chapa
haha apartamento professor
tá_a_fim tattoo ir_pra_aula campus faculdade
página
na do_caralho estudar na_biblioteca
fodido
cansei estudando
este_artigo dormitório nãoposso
bêbado
sem quê semestre
hahaha fudeu
palestra sou
prova_final aula puta_merda
entrevista
semestral última colegas
prova lab
álcool
destruído
bebendo
bêbado
bebaço
bar beber
ressaca
fudido
bebaço sóbrio
chapado
um_pouco aguenta
zureta
bosta
fudido merda
cara
vadias bêbaço foda
mulherada do_caralho
merda
putz
porra foda-se bunda
mentira
terminou prazo
provas aulas final
acabou tarefas
artigos projetos
dever_de_casa semana semanas
semestre testes escrevendo

23 a 29 anos

repetir
não
uni
jardim
dias
trabalho
horário
voluntário
colocação
alugar
mudar
financiar
complexo
colega
quarto
apartamento
mês
mudando
casa
área
assinei
lugar
apt
interessado
cargo
entrevista
contratar
vendas
interessado
currículo
carreira
assistente
gerente
experiência
emprego
escritório
negócios
contratando
empresa
divertido
gratidão
casado
mercado
avisa
abraços
empolgado
cuidar_de
uma_bebida
novo_emprego
bebendo
curtindo
o_bar
empresa
todos
apartamento
dias_para
:-(
impostos
novo_lugar
turno
no_trabalho
mulheres
@
marido
trabalhar_amanhã
pronto
apt
relaxando
coitadinha
casamento
.
folga
lavanderia
contas
cervejas
cerveja
abençoado
vinho
tarefas
escritório
necessário
meus_bebês
depois_trabalho
marido
para_curtir
drinks
férias
celebrando
dedos_cruzados
celebrar
contos
impostos
conta
mês
dívida
cheque
aluguel
contas
devo
dinheiro
paguei
emprego
pagar
pagando
salário
cervejas
raízes
bebida
bebendo
caneca
latas
chopp
ginger
garrafas
gelada
cerveja
bar
pong
beber
indo
cedo
casa
maridão
descansar
jantar
relaxar
cochilo
cama
pronto
trabalhar
relaxando
banho
comer

30 a 65 anos

Chamo esse gráfico de "Beber. Trabalhar. Rezar." Na adolescência as pessoas bebem. Na faixa dos vinte anos, trabalham. Dos trinta em diante, rezam.

Uma nova ferramenta poderosa para avaliar textos é a chamada análise de estado emocional. Cientistas podem hoje estimar o grau de felicidade ou tristeza em um determinado trecho de texto.

Como? Equipes de cientistas pediram a um grande número de pessoas para codificar dezenas de milhares de palavras em inglês como positivas ou negativas. As palavras mais positivas, de acordo com esta metodologia, incluem "felicidade", "amor" e "maravilhoso". As mais negativas, "triste", "morte" e "depressão". Então, eles criaram um índice de humor com um gigantesco banco de palavras.

Utilizando este índice, eles conseguiram medir o humor médio em um fragmento de texto. Se alguém escreve: "Estou feliz e apaixonado e me sentindo maravilhoso", a análise de estado emocional classificaria o texto como extremamente feliz. Se alguém escreve: "Estou triste pensando em toda a morte e depressão do mundo", a análise de

estado emocional classificaria o trecho como um texto extremamente triste. Outros trechos de texto estariam em algum ponto entre eles.

Assim, o que podemos descobrir ao codificar o humor de um texto? Cientistas de dados do Facebook mostraram uma empolgante possibilidade. Eles conseguiram estimar a Felicidade Nacional Bruta de um país. Se as mensagens de status tendem a ser positivas, o país é considerado feliz naquele dia. Se tendem a ser negativas, o país é considerado triste naquele dia.

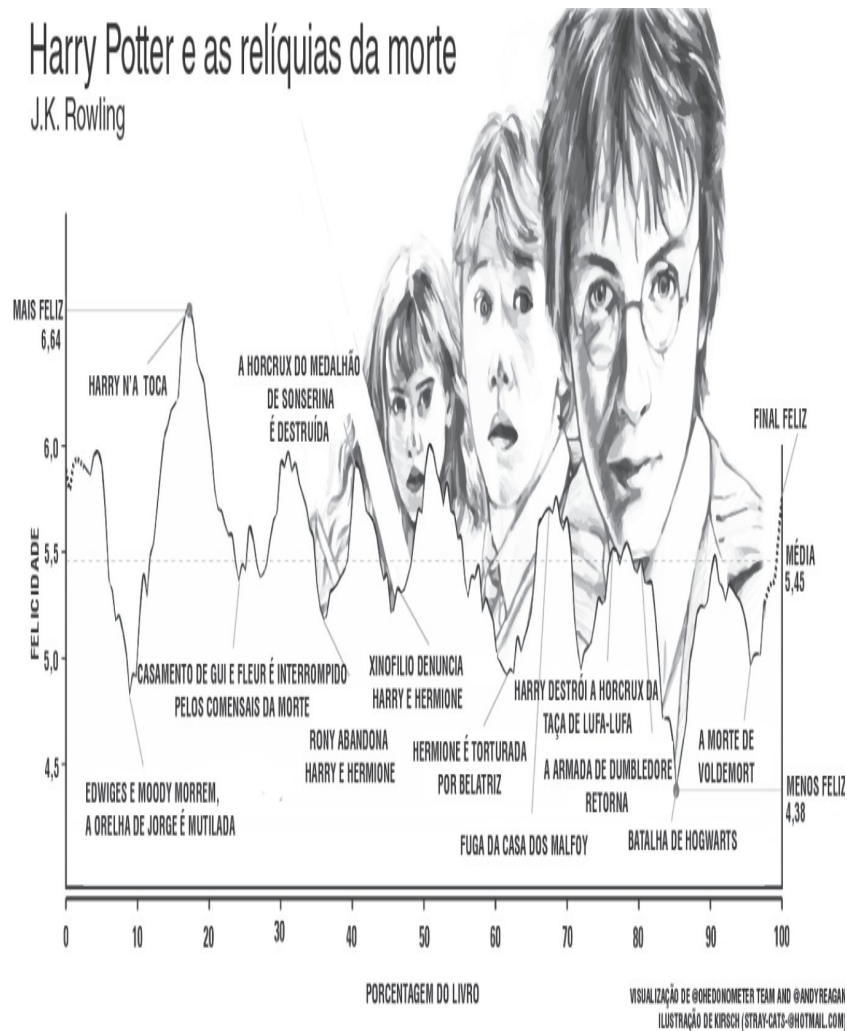
Entre as descobertas dos cientistas de dados do Facebook estão: o Natal é um dos dias mais felizes do ano. Mas eu estava um tanto cético sobre esta análise — e um tanto cético sobre todo o projeto. Normalmente, penso que muitas pessoas escondem que, na verdade, ficam tristes no Natal, porque estão solitárias ou pelos conflitos familiares. De modo mais genérico, tendo a não confiar nas atualizações de status do Facebook, por motivos que discutiremos neste capítulo — especialmente nossa propensão a mentir sobre nossas vidas na mídia social.

Se você está sozinho e infeliz no Natal, realmente quer incomodar todos seus amigos postando sobre o quanto está infeliz? Desconfio que muitas pessoas passando um Natal infeliz ainda postem mensagens no Facebook sobre o quanto estão gratas por sua “vida maravilhosa, fantástica, feliz e fabulosa”. Essas mensagens são então codificadas, aumentando substancialmente a Felicidade Nacional Bruta dos Estados Unidos. Se quisermos realmente codificar a Felicidade Nacional Bruta, temos que usar mais fontes do que apenas as atualizações de status do Facebook.

Dito isso, a descoberta de que o Natal é, em média, uma ocasião feliz parece legitimamente verdadeira. As buscas no Google por depressão e as pesquisas do Gallup também indicam que o Natal é um dos dias mais felizes do ano. E, contrariando a lenda urbana, os suicídios diminuem nas festas de fim do ano. Mesmo havendo pessoas tristes e solitárias no Natal, há mais pessoas felizes.

Hoje em dia, quando as pessoas se sentam para ler, a maior parte do tempo é gasta analisando atualizações de status do Facebook. Mas, era uma vez, não há muito tempo, que as pessoas liam histórias, às vezes em livros. A análise de estado emocional tem muito a nos ensinar sobre isso.

Uma equipe de cientistas liderada por Andy Reagan, hoje na Berkeley School of Information da Universidade da Califórnia, baixou textos contidos em milhares de livros e roteiros de filmes. Eles então codificaram o grau de felicidade ou tristeza de cada ponto da história.

Por exemplo, no livro *Harry Potter e as Relíquias da Morte*, veja a seguir como o clima da história muda, de acordo com a equipe de cientistas, juntamente com uma descrição dos pontos-chave da trama.

Observe que os altos e baixos no clima detectados pela análise de estado emocional correspondem a eventos-chave na trama.

A maioria das histórias tem estruturas mais simples. Analise, por exemplo, a tragédia de Shakespeare *Rei João*. Nesta peça, nada dá certo. O Rei João da Inglaterra é forçado a renunciar ao trono. Ele é excomungado por desobedecer ao papa. A guerra é deflagrada. Seu sobrinho morre, suspeita-se de suicídio. Outras pessoas morrem. Finalmente, João é envenenado por um monge descontente.

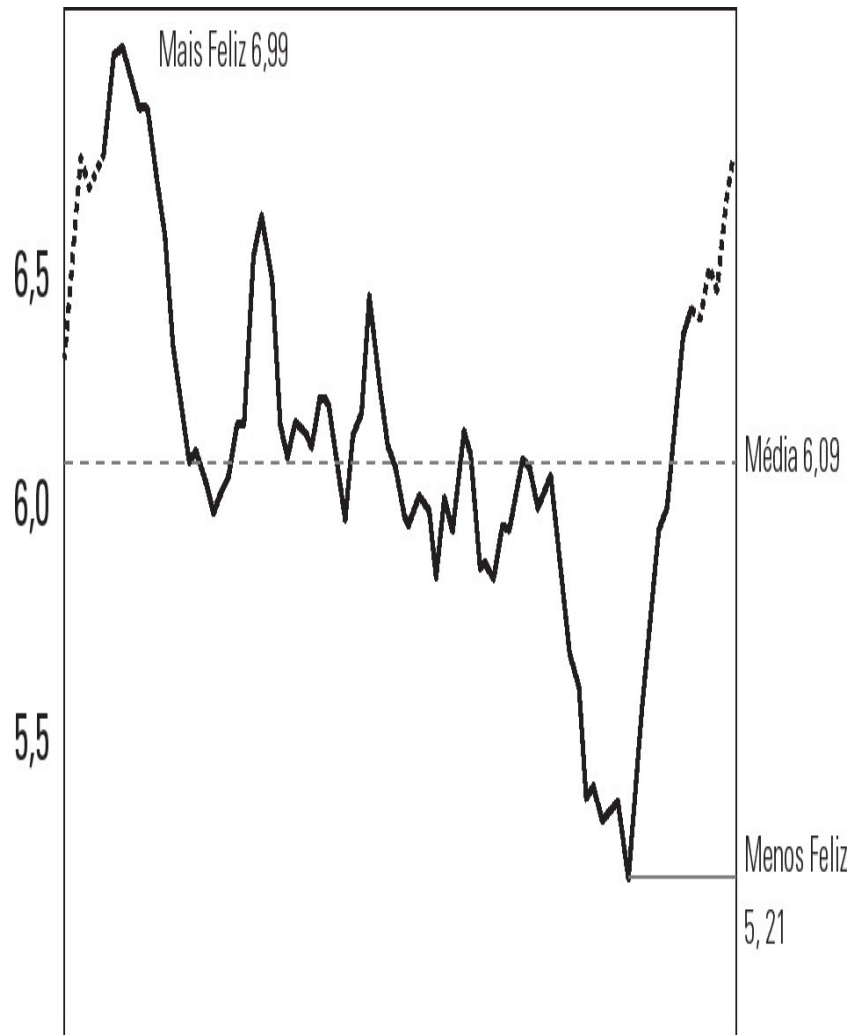
A seguir, a análise de estado emocional durante a evolução da peça.

Em outras palavras, apenas pela análise das palavras, o computador foi capaz de detectar que as coisas vão de mal a pior.

Observe o filme *127 Horas*. A trama resumida do filme é:

> Um montanhista visita o Parque Nacional de Cânions em Utah para escalar. Ele conhece outros trilheiros, mas se separa deles. De repente, escorrega e despenca em uma fenda, ficando com a mão e o punho presos em uma rocha. Ele tenta várias maneiras de se soltar, sem sucesso. Seu ânimo se esvai e ele fica bastante deprimido. Finalmente, ele amputa o braço para escapar. Ele se casa, começa uma família e continua escalando, mas agora passa a deixar um bilhete informando sempre onde está.

A seguir mostro a análise de estado emocional da evolução do filme, feita pela equipe de cientistas de Reagan.

127 Horas direção de Danny Boyle

Assim, o que aprendemos analisando o astral de milhares dessas histórias?

Os cientistas de computação descobriram que um grande porcentual das histórias se encaixa em seis estruturas relativamente simples. Usando o gráfico da equipe de Reagan, temos:

Da Pobreza para a Riqueza (ascensão)
Riqueza para Pobreza (queda)
Homem em uma Enrascada (queda, depois ascensão)
Ícaro (ascensão, depois queda)
Cinderela (ascensão, depois queda, depois ascensão)
Édipo (queda, depois ascensão, depois queda)

Pode haver pequenas guinadas não captadas por este esquema simples. Por exemplo, *127 Horas* é uma história do tipo Homem em uma Enrascada, apesar de haver momentos durante a queda em que o estado emocional temporariamente se eleva. A estrutura geral e abrangente da maioria das histórias se encaixa em uma das seis categorias. *Harry Potter e as Relíquias da Morte* é uma exceção.

Há muitas outras perguntas que podemos responder. Por exemplo, como a estrutura das histórias mudou ao longo do tempo? As histórias se tornaram mais complexas no decorrer dos anos? As culturas diferem nos tipos de histórias que contam? De que tipos de história as pessoas mais gostam? Homens e mulheres são atraídos por diferentes estruturas narrativas? E pessoas de países diferentes?

Por fim, o uso de texto como dado nos dá percepções inéditas sobre o que o público quer de verdade, que pode ser diferente do que autores ou produtores pensam que ele quer. Já temos pistas apontando nesta direção.

Observe o estudo de dois professores da Wharton School, Jonah Berger e Katherine L. Milkman, sobre quais tipos de histórias são compartilhadas. Eles testaram se histórias positivas ou negativas tinham maior probabilidade

de chegar à lista de notícias mais compartilhadas por e-mail do *New York Times*. Eles baixaram todos os artigos do *Times* durante um período de três meses. Usando a análise de estado emocional, os professores codificaram o astral dos artigos. Exemplos de histórias classificadas como positivas foram: “Recém-chegados Se Apaixonam pela Cidade” e “Prêmio Tony para Filantropia”. Histórias como “Boatos na Internet Seriam a Causa de Suicídio de Atriz Coreana” e “Alemanha: Tratador de Filhote de Urso-polar Morre” foram consideradas, como era de se esperar, negativas.

Os professores também tinham informação sobre onde a história aparecia. Estava em uma página inicial? Do lado superior direito? Do superior esquerdo? E tinham informações sobre quando a história foi publicada. Quinta-feira à noite? Segunda de manhã?

Eles puderam comparar dois artigos — um positivo e um negativo — que apareceram em um local semelhante no site do *Times* e que foram divulgados em um momento parecido, e ver qual deles tinha maior probabilidade de ser encaminhado por e-mail.

Então, qual deles é compartilhado, o positivo ou o negativo?

Os positivos. De acordo com a conclusão dos autores: “Quanto mais positivo o conteúdo maior a probabilidade de viralizar.”

Note que isso parece contrastar com a sabedoria jornalística tradicional de que as pessoas são atraídas por histórias violentas e trágicas. É bem verdade que a mídia de notícias oferece uma profusão de histórias sombrias ao público. Há algo de verdadeiro no ditado jornalístico que diz: “Se sangra, anda.” O estudo dos professores da Wharton, porém, indica que as pessoas de fato preferem histórias mais alegres e positivas. Pode até sugerir um novo ditado: “Se é feliz, é compartilhado”, apesar da falta de rima.

Chega de falar de texto feliz e triste. Como podemos descobrir se um texto é liberal ou conservador? E o que isso nos mostra sobre a mídia de notícias moderna? Isto é um pouco mais complicado, e nos leva de volta para Gentzkow e Shapiro. Lembra deles? Os economistas que perceberam a diferença na forma de descrever o casamento gay em dois jornais distintos e imaginaram se conseguiriam usar a linguagem para revelar inclinações políticas.

A primeira coisa que esses dois jovens acadêmicos ambiciosos fizeram foi examinar as transcrições dos *Registros do Congresso*. Como este registro já estava digitalizado, eles coletaram todas as palavras utilizadas por cada congressista Democrata e Republicano em 2005. Então verificaram se determinadas frases eram significativamente mais usadas por Democratas ou Republicanos.

Algumas de fato foram. Veja exemplos de cada categoria.

FRASES USADAS MUITO MAIS POR DEMOCRATAS	FRASES USADAS MUITO MAIS POR REPUBLICANOS
Imposto imobiliário	Imposto “de morte”
Privatizar a Previdência Social	Reformar a Previdência Social
Rosa Parks	Saddam Hussein
Direitos trabalhistas	Direitos sobre propriedade privada
Pessoas pobres	Gastos do governo

O que explica essas diferenças na linguagem?

Às vezes Democratas e Republicanos usam expressões diferentes para descrever o mesmo conceito. Em 2005, os Republicanos tentaram extinguir o imposto federal de transmissão sucessória. Eles costumam descrevê-lo como “imposto ‘de morte’” (que soa como uma imposição sobre o recém-falecido). Os Democratas o descrevem como “imposto imobiliário” (que faz parecer que é um imposto sobre a riqueza). De modo semelhante, os Republicanos tentaram transformar a Previdência Social em contas individuais de aposentadoria. Para os Republicanos, isto era uma “reforma”. Para os Democratas, algo aparentemente mais perigoso, “privatização”.

Às vezes as diferenças na linguagem são uma questão de ênfase. Republicanos e Democratas presumivelmente têm grande respeito por Rosa Parks, a heroína dos direitos civis. Mas os Democratas falam sobre ela com mais

frequência. Do mesmo modo, Democratas e Republicanos, obviamente, acham que Saddan Hussein, ex-líder do Iraque, era um ditador cruel. Mas os Republicanos citaram seu nome repetidas vezes para justificar a Guerra do Iraque. Do mesmo modo, os "direitos trabalhistas" e a preocupação com "pessoas pobres" são princípios essenciais do Partido Democrata. "Direitos de propriedade privada" e redução de "gastos do governo" são princípios essenciais dos Republicanos.

E essas diferenças no uso da linguagem são substanciais. Por exemplo, em 2005, congressistas Republicanos utilizaram a expressão "imposto de 'morte'" 365 vezes e "imposto imobiliário" apenas 46 vezes. Para os congressistas Democratas, o padrão foi o inverso. Eles usaram a expressão "imposto 'de morte'" apenas 35 vezes e "imposto imobiliário" 195 vezes.

Se essas palavras podem nos dizer se um congressista é Democrata ou Republicano, os acadêmicos observaram, poderiam também nos dizer se um jornal tem inclinação política para a esquerda ou para a direita. Assim como um congressista Republicano é mais propenso a usar a expressão "imposto 'de morte'" para incutir uma resposta negativa do público, jornais conservadores podem fazer o mesmo. O jornal relativamente liberal *Washington Post* usou a expressão "imposto imobiliário" com uma frequência 13,7 vezes maior do que "imposto 'de morte'". O conservador *Washington Times* usou "imposto 'de morte'" e "imposto imobiliário" aproximadamente com a mesma frequência.

Graças às maravilhas da internet, Gentzkow e Shapiro puderam analisar o uso da linguagem em um grande número de jornais dos Estados Unidos. Eles utilizaram dois sites, *newslibrary.com* e *proquest.com*, que juntos digitalizaram 433 jornais. Então, contaram com que frequência mil dessas expressões carregadas de significado político foram usadas para mensurar a inclinação política de um jornal. O jornal mais liberal, de acordo com essa avaliação, foi o *Philadelphia Daily News;* o mais conservador: o *Billings Gazette* (de Montana).

Quando se tem a primeira mensuração abrangente da tendência da mídia para uma ampla gama de informações, talvez seja possível responder a pergunta mais importante sobre a imprensa: por que algumas publicações tendem para a esquerda e outras, para a direita?

Os economistas rapidamente identificaram um fator-chave: a política de determinada área. Se a região é amplamente liberal, como a Filadélfia e Detroit, o principal jornal tende a ser liberal. Se é mais conservadora, como Billings e Amarillo, Texas, o jornal dominante tende a ser conservador. Em outras palavras, as provas sugerem fortemente que os jornais tendem a oferecer o que os leitores querem.

Você poderia pensar que o dono do jornal teria alguma influência na tendência de sua cobertura jornalística, mas, via de regra, o dono do jornal tem menos influência em sua tendência política do que imaginávamos. Observe o que acontece quando a mesma pessoa ou empresa é dona de jornais em mercados diferentes. Pense na empresa *The New York Times.* Ela é dona do liberal, segundo os resultados de Gentzkow e Shapiro, *New York Times*, na cidade de Nova York, onde aproximadamente 70% da população é Democrata. Ela também era dona, na época do estudo, do conservador, de acordo com a avaliação, *Spartanburg Herald-Journal,* em Spartanburg, Carolina do Sul, onde aproximadamente 70% da população é Republicana. Há exceções, é claro: a Rupert Murdoch's News Corporation é dona do *New York Post*, considerado conservador por praticamente todo mundo. Mas, de modo geral, as descobertas sugerem que é o mercado, muito mais do que os donos, que determinam as tendências de um jornal.

O estudo teve um profundo impacto em como pensamos sobre a mídia de notícias. Muitas pessoas, especialmente marxistas, consideravam o jornalismo norte-americano controlado por pessoas ou empresas ricas com o objetivo de influenciar as massas, talvez incutindo nas pessoas suas visões políticas. O artigo de Gentzkow e Shapiro indica, porém, que essa não é a motivação predominante dos proprietários. Em vez disso, os controladores da imprensa norte-americana oferecem primordialmente ao público o que ele quer, para que os donos fiquem cada vez mais ricos.

Ah, mais uma pergunta — importante, controversa e até polêmica. A mídia de notícias dos Estados Unidos, em média, tende à direita ou à esquerda? Na média, é conservadora ou liberal?

Gentzkow e Shapiro descobriram que jornais tendem à esquerda. O jornal mediano é mais semelhante, nas palavras que usa, a um congressista Democrata do que a um Republicano.

"Ahá! Eu avisei!", gritam os leitores conservadores. Muitos conservadores há muito tempo desconfiam que os jornais sempre foram tendenciosos para tentar manipular o povo a apoiar seus pontos de vista de esquerda.

Nem tanto, dizem os autores. De fato, a tendência liberal é bem dosada para aquilo que os leitores querem. Leitores de jornais, em média, são de esquerda. (Eles têm dados sobre isso.) E os jornais, em geral, tendem um pouco à esquerda para atender à perspectiva de seus leitores.

Não existe conspiração. É apenas capitalismo.

A mídia de notícias, sugerem os resultados de Gentzkow e Shapiro, normalmente opera como qualquer outra indústria no planeta. Assim como os supermercados lotam suas prateleiras com o sorvete que as pessoas querem, os jornais descobrem qual é o ponto de vista buscado por seus leitores e enchem suas páginas com ele. "São só

negócios”, disse-me Shapiro. É isso que podemos aprender ao discriminar e quantificar questões intricadas como as notícias, análises e opiniões em seus elementos constitutivos: as palavras.

FOTOS COMO DADOS

Tradicionalmente, quando acadêmicos ou empresários queriam dados, realizavam pesquisas. Os dados vinham cuidadosamente organizados, extraídos de números ou alternativas assinaladas em questionários. Não é mais assim. Os dias dos dados baseados em pesquisas, estruturados, limpos e simples chegaram ao fim. Nesta nova era, os rastros confusos que deixamos ao longo da vida estão se tornando a principal fonte de dados.

Como já vimos, palavras são dados. Cliques são dados. Links são dados. Erros de digitação são dados. Bananas em sonhos são dados. O tom de voz é um dado. Assobios na respiração são dados. Ritmo cardíaco é dado. Tamanho do baço é dado. As buscas são, é o que defendo, os dados mais reveladores.

Fotos também são dados.

Assim como as palavras, que antes eram confinadas em livros e periódicos em prateleiras empoeiradas, agora foram digitalizadas, as fotos foram libertas de álbuns e caixas de papelão. Elas também foram transformadas em bits e armazenadas em nuvem.

E assim como o texto — que nos mostra, por exemplo, a mudança nas formas que as pessoas falavam —, as fotos nos proporcionam lições de história — revelando, por exemplo, as mudanças na forma que as pessoas posavam para as fotos.

Observe este engenhoso estudo feito por uma equipe de quatro cientistas de computação em Brown e Berkeley. Utilizando um excelente desenvolvimento da era digital: muitas escolas de ensino médio escanearam seus anuários e os disponibilizaram online. Em toda a internet, os pesquisadores encontraram 949 anuários escaneados de escolas de ensino médio dos Estados Unidos entre os anos de 1905 e 2013. Isto inclui dezenas de milhares de retratos dos formandos. Usando um software, eles foram capazes de criar um rosto “médio” para cada década. Em outras palavras, conseguiram descobrir a localização média e os traços de narizes, olhos, lábios e cabelos das pessoas. Veja a seguir os rostos médios ao longo de mais de um século, separados por gênero:

Percebeu algo? Os norte-americanos — e especialmente as mulheres — começaram a sorrir. Os rostos mudaram de sérios no início do século XX para sorridentes no final.

Por que a mudança? Os norte-americanos ficaram mais felizes?

Não. Outros acadêmicos ajudaram a responder esta pergunta. O motivo é, pelo menos para mim, fascinante. Quando a fotografia foi inventada, as pessoas pensavam nelas como quadros. Não havia algo que pudesse servir de comparação. Assim, pessoas fotografadas copiavam as pessoas nos quadros. E como nos quadros era impossível manter um sorriso pelas muitas horas que levavam para ficar prontos, as pessoas adotavam uma fisionomia séria. E assim, ao serem fotografadas, faziam o mesmo.

O que finalmente as fez mudar? Negócios, lucro e marketing, é claro. Em meados do século XX, a Kodak, a empresa de filmes e câmeras, estava frustrada pelo pequeno número de fotos que as pessoas faziam, e criou uma estratégia para aumentar esse número. As propagandas da Kodak começaram a associar fotos com felicidade. O objetivo era fazer com que as pessoas se habituassem a fazer fotos sempre que quisessem mostrar para os outros como estavam se divertindo. Todas essas fotos de pessoas sorridentes nos anuários são o resultado de uma campanha de sucesso (assim como a maioria das fotos que vemos no Facebook e no Instagram hoje).

Mas fotos como dados podem nos mostrar muito mais do que formandos do ensino médio começando a dizer “X”. Surpreendentemente, as fotos podem nos dizer como está a economia.

Analisemos um artigo acadêmico provocativamente intitulado “Medindo o Crescimento Econômico a partir do Espaço Sideral”. Quando um artigo tem um nome como este, pode apostar que vou ler. Os autores — J. Vernon Henderson, Adam Storeygard e David N. Weil — começaram a perceber que em muitos países em desenvolvimento, as medições do PIB nacional são ineficientes. Isso se deve ao fato de grande parte da atividade econômica não ser contabilizada, e os órgãos governamentais destinados a medir os resultados econômicos terem recursos limitados.

Qual a ideia não convencional dos autores? Eles poderiam medir o PIB com base na quantidade de luzes acesas à noite nesses países. Utilizando a informação das fotos feitas por um satélite militar da Força Aérea dos Estados Unidos que circula a terra catorze vezes por dia.

Por que as luzes acesas à noite seriam um bom medidor do PIB? Bem, em partes muito pobres do mundo, as pessoas têm dificuldades para pagar por energia elétrica. E como resultado, quando a situação econômica está ruim, as casas e vilarejos reduzem dramaticamente a quantidade de luzes que acendem à noite.

A iluminação noturna diminuiu drasticamente na Indonésia durante a crise asiática de 1998. Na Coreia do Sul, a iluminação noturna aumentou 72% entre 1992 e 2008, correspondendo a um notável forte crescimento econômico ao longo deste período. Na Coreia do Norte, durante o mesmo período, a iluminação noturna na verdade diminuiu, equivalendo ao desempenho sombrio da economia neste período.

Em 1998, no sul de Madagascar, uma grande reserva de rubis e safiras foi descoberta. A cidade de Ilakaka foi de um pouco mais de uma parada de caminhões para um grande centro comercial. Não havia praticamente luz alguma em Ilakaka à noite antes de 1998. Nos próximos cinco anos, houve uma explosão de luz à noite.

Os autores admitem que os dados sobre a iluminação noturna estão longe de ser uma medição perfeita do desempenho econômico. Definitivamente não é possível saber exatamente a situação econômica apenas pela quantidade de luz captada por um satélite à noite. Eles não recomendam de forma alguma essa medição para países desenvolvidos, como os Estados Unidos, onde existem dados econômicos muito mais precisos. E, na verdade, mesmo em países em desenvolvimento, os autores acham que a iluminação noturna é somente tão útil quanto as medições oficiais. Mas a combinação dos dados falhos do governo e dos dados imperfeitos das luzes noturnas nos dá uma estimativa melhor do que apenas um deles sozinho. Em outras palavras, é possível aprimorar seu entendimento sobre as economias em desenvolvimento usando fotos feitas a partir do espaço.

Joseph Reisinger, doutor em ciência de computação, compartilha da frustração dos autores do artigo sobre os conjuntos de dados existentes referentes à economia dos países em desenvolvimento. Em abril de 2014, observa Reisinger, a Nigéria atualizou sua estimativa de PIB, levando em consideração novos setores que haviam ignorado anteriormente. A estimativa do PIB foi 90% maior.

“A Nigéria é a maior economia da África”, disse Reisinger. “Não sabemos sequer a questão mais básica que gostaríamos de saber sobre aquele país.”

Ele queria encontrar uma maneira de ter uma visão mais apurada sobre o desempenho econômico. Sua solução é um belo exemplo de como reinventar o que consideramos dados e seu valor.

Reisinger descobriu uma empresa, Premise, que emprega um grupo de funcionários em países em desenvolvimento, munidos de smartphones. O trabalho deles? Fazer fotos de fatos interessantes com relevância econômica.

Os funcionários podiam fazer fotos em postos de gasolina ou em bancas de frutas em supermercados. Faziam fotos do mesmo lugar diversas vezes. As imagens eram enviadas para a Premise, e um segundo grupo de funcionários — cientistas de computação — as transformava em dados. Os analistas da empresa podiam codificar tudo, desde o tamanho das filas nos postos de combustíveis; quantas maçãs têm na banca do supermercado; a qualidade das maçãs à venda, até o preço da caixa de maçãs no mercado. Com base nas fotos de todos os tipos de atividade, a Premise é capaz de fornecer estimativas de resultados econômicos e da inflação. Em países em desenvolvimento, longas filas em postos de combustíveis são o principal indicador de problemas econômicos. Assim como a indisponibilidade ou maçãs colhidas antes do amadurecimento. As fotos feitas pela Premise na China ajudaram a descobrir a inflação de alimentos em 2011 e a deflação de alimentos em 2012, muito antes dos dados oficiais.

A Premise vende essas informações para bancos ou fundos hedge e ainda colabora com o Banco Mundial.

Assim como muitas boas ideias, a da Premise é um presente que continua rendendo frutos. O Banco Mundial recentemente se interessou pelo tamanho da economia clandestina de cigarros nas Filipinas. Eles queriam saber, especialmente, os efeitos dos recentes esforços do governo, que incluíram blitzes, para pegar fabricantes que não pagavam impostos. A ideia engenhosa da Premise? Fazer fotos de maços de cigarros vistos nas ruas. Verificar quantos deles tinham selos de impostos, como todo cigarro legítimo deve ter. Eles descobriram que essa parte da economia clandestina, que era grande em 2015, reduziu significativamente em 2016. Os esforços do governo funcionaram, embora conseguir perceber algo normalmente tão escondido — cigarros ilegais — tenha exigido novos dados.

Como vimos, o que são dados foi amplamente reinventado na era digital, e muitas novas percepções têm sido descobertas com essas novas informações. Aprender o que provoca a tendência da mídia, o que torna um primeiro encontro bom e qual a situação de economias em desenvolvimento é apenas o início.

Não por acaso, muito dinheiro é gerado a partir desses novos dados, começando com as dezenas de bilhões de dólares de Brin e Page. Joseph Reisinger também não se saiu mal. Observadores estimam que a Premise hoje lucra

dezenas de milhões de dólares anualmente. Investidores recentemente aplicaram US$50 milhões na empresa. Isto significa que alguns investidores consideram a Premise um dos empreendimentos mais valiosos do mundo, com a atividade principal de fazer e vender fotos, no mesmo patamar que a *Playboy*.

Em outras palavras, existe um gigantesco valor, para acadêmicos e empreendedores, na utilização de todos os novos tipos de dados hoje disponíveis, em pensar de modo mais abrangente sobre o que são dados. Atualmente, um cientista de dados não precisa se limitar a uma visão estreita e tradicional dos dados. A abundância de produtos nas gôndolas do supermercado são dados. A qualidade das maçãs são dados. As fotos feitas do espaço são dados. As curvaturas dos lábios são dados. Tudo é dado!

E com todos esses novos dados, podemos finalmente enxergar algo além do que todo mundo mente.

* Embora as versões iniciais do Google Flu tivessem falhas significativas, os pesquisadores recentemente recalibraram o modelo, com mais sucesso.

* Em 1998, ao digitar “carros” em um mecanismo de busca pré-Google, você receberia uma enxurrada de sites pornográficos. Esses sites de pornografia escreviam a palavra “carros” geralmente em branco sobre o fundo branco para enganar o mecanismo de busca. Assim conseguiam mais cliques de pessoas que, na verdade, queriam comprar um carro, mas que acabavam se distraindo com a pornografia.

* Uma teoria em que estou trabalhando: Big Data apenas confirma tudo que o falecido Leonard Cohen sempre dizia. Por exemplo, ele certa vez deu ao sobrinho o seguinte conselho para conquistar uma mulher: “Ouça muito. E depois ouça um pouco mais.“ Isto parece bem semelhante ao que estes cientistas descobriram.

4

SORO DIGITAL DA VERDADE

Todo mundo mente.

As pessoas mentem sobre o quanto beberam antes de chegar em casa. Mentem sobre a frequência com que vão à academia, sobre quanto custou o sapato novo ou se leram determinado livro. Elas faltam ao trabalho por doenças fictícias. Dizem que manterão contato e desaparecem. Falam que o problema não é você, quando é. Declaram um amor que não sentem. Proclamam aos quatro ventos sua felicidade, quando estão em frangalhos. Alegam gostar de mulheres quando, de fato, preferem homens.

As pessoas mentem para os amigos. Mentem para os chefes. Para os filhos. Para os pais. Mentem para os médicos. Para maridos e esposas. Mentem para si mesmas.

E, com certeza, mentem para pesquisas.

Vamos fazer uma pequena pesquisa:

Você já colou em uma prova? ___________
Já fantasiou sobre matar alguém? __________

Ficou tentado a mentir? Muitas pessoas omitem seus comportamentos e pensamentos desagradáveis em pesquisas. Elas querem passar uma boa imagem, mesmo quando o anonimato é garantido. Isto se chama propensão de desejabilidade social.

Um importante artigo de 1950 forneceu sólidas evidências de como as pesquisas podem padecer em razão desta propensão. Os pesquisadores coletaram dados, de fontes oficiais, sobre os residentes de Denver: que percentual deles votou, fez doações para caridade e possuía um cartão de biblioteca. Então, entrevistaram os moradores para checar se os números batiam. Os resultados foram, na época, chocantes. As informações fornecidas na pesquisa foram muito diferentes dos dados coletados pelos pesquisadores. Apesar de ninguém se identificar, as pessoas, em grande proporção, exageram nas informações de registro como eleitor, comparecimento na eleição e doação para caridade.

	RELATADO NA PESQUISA	DADOS OFICIAIS
Eleitores registrados para votar	83%	69%
Votaram na última eleição presidencial	73%	61%
Votaram na última eleição para prefeito	63%	36%
Possuem um cartão de biblioteca	20%	13%
Doaram para uma campanha de caridade recente	67%	33%

Algo mudou em 65 anos? Na era da internet, não possuir um cartão de biblioteca não é mais embaraçoso. Mas, apesar do que pode ser considerado desejável ou vergonhoso ter mudado, a tendência das pessoas de enganar os pesquisadores continua a mesma.

Uma pesquisa recente questionou os graduandos da Universidade de Maryland sobre suas experiências na universidade. As respostas foram comparadas aos registros oficiais. As pessoas deram informações substancialmente erradas, de forma que passassem uma boa imagem de si mesmas. Menos de 2% relatou ter se formado com uma média inferior a 2,5. (Na realidade, aproximadamente 11% delas obtiveram tal média.) E 44% disseram ter doado dinheiro para a universidade no último ano. (Na realidade, o percentual foi de 28%.)

Certamente é possível que a mentira tenha desempenhado um importante papel no fracasso das pesquisas de intenção em prever a vitória de Donald Trump nas eleições de 2016. As pesquisas eleitorais, em média, subestimaram o apoio a Trump em cerca de dois pontos percentuais. Algumas pessoas podem ter ficado envergonhadas em declarar seu apoio. Algumas alegaram indecisão quando de fato já pretendiam votar em Trump.

Por que as pessoas omitem informações em pesquisas anônimas? Perguntei a Roger Tourangeau, pesquisador emérito da Universidade de Michigan e talvez o maior especialista do mundo em desvio de desejabilidade social. “Nossa inclinação para as ‘mentiras brancas’ é uma importante parte do problema”, explicou Tourangeau. “Cerca de um terço do tempo, as pessoas mentem na vida real. Este hábito é transferido para as pesquisas.”

Além disto, às vezes temos o estranho hábito de mentir para nós mesmos. “Existe uma relutância em admitir para si mesmo que, digamos, fomos um fracasso como alunos”, diz Tourangeau.

Mentir para si mesmo pode explicar por que tantas pessoas dizem estar acima da média. Qual é o tamanho deste problema? Mais de 40% dos engenheiros de uma empresa disseram que estavam entre os 5% melhores. Mais de 90% dos professores universitários dizem fazer um trabalho acima da média. Um quarto dos alunos do último ano do ensino médio acredita estar entre o 1% das pessoas com maior habilidade social. Se está iludindo a si mesmo, não será possível ser honesto em uma pesquisa.

Outro fator determinante para mentirmos em pesquisas é nosso forte desejo de criar uma boa impressão para o estranho que a conduz, quer dizer, se houver alguém a conduzindo. Como diz Tourangeau: “Uma pessoa que parece com sua tia preferida se aproxima… Você quer contar para sua tia preferida que fumou maconha no último mês?”* Gostaria de admitir que não doou dinheiro para sua querida instituição de ensino?

Por esta razão, quanto mais impessoais forem as condições, mais honestas as pessoas serão. Para obter respostas verdadeiras, as pesquisas da internet são melhores do que as por telefone, que por sua vez são melhores do que pessoais. As pessoas serão mais francas se estiverem sozinhas do que se houver outras pessoas na sala.

Entretanto, em assuntos delicados, todos os métodos de pesquisa evocarão uma quantidade substancial de informação incorreta. Tourangeau usou aqui uma palavra que é frequentemente empregada por economistas: incentivo. As pessoas não têm qualquer incentivo para dizer a verdade.

Como, então, podemos aprender o que nossos colegas seres humanos estão realmente pensando ou fazendo?

Em alguns casos, existem fontes de dados oficiais que podemos consultar para obter a verdade. Mesmo que as pessoas mintam sobre as doações para caridade, por exemplo, podemos conseguir os números reais de doações em determinada região das próprias entidades. Mas quando tentamos conhecer comportamentos que não são listados em registros oficiais ou descobrir o que as pessoas pensam — suas verdadeiras crenças, sentimentos e desejos —, não há outra fonte de informação, exceto o que as pessoas se permitem dizer nas pesquisas. Isto é, até agora.

Esse é o segundo poder do Big Data: determinadas fontes online conseguem fazer com que as pessoas admitam coisas que não admitiriam em nenhum outro lugar. Elas agem como um soro digital da verdade. Pense nas buscas do Google. Lembre-se das condições que tornam as pessoas mais honestas. Online? Sim. Sozinha? Sim. Ninguém conduzindo a pesquisa? Sim.

E há outra imensa vantagem que as buscas no Google têm em fazer as pessoas dizerem a verdade: incentivos. Se você gosta de piadas racistas, tem zero incentivo de compartilhar este fato politicamente incorreto em uma pesquisa. No entanto, existe um incentivo para buscar novas piadas racistas online. Se pensa que pode estar sofrendo de depressão, não tem incentivo algum para admitir isto para uma pesquisa. Mas sim para perguntar ao Google sobre os sintomas e potenciais tratamentos.

Mesmo que esteja mentindo para si mesmo, ainda assim o Google pode saber a verdade. Alguns dias antes da eleição, você e seus vizinhos poderiam legitimamente achar que compareceriam à seção de votação para votar. Mas se você ou eles não buscaram qualquer informação sobre como ou onde votar, cientistas de dados como eu podem estimar que a taxa de comparecimento em sua região será baixa. Da mesma forma, talvez você não tenha admitido sequer para si mesmo que sofre de depressão, mesmo fazendo buscas no Google sobre crises de choro e dificuldade de sair da cama. Entretanto, você apareceria nas buscas relacionadas à depressão de determinada região, analisadas anteriormente neste livro.

Pense na própria experiência usando o Google. Aposto que você já deve ter digitado coisas na caixa de busca que revelam um comportamento ou pensamento que hesitaria em admitir em uma situação formal. Na verdade, há uma torrente de provas de que a grande maioria dos norte-americanos conta ao Google coisas muito pessoais. Por exemplo, eles buscam mais pela palavra “pornô” do que por “clima”. Isso é difícil, aliás, de conciliar com os dados de pesquisa, já que apenas 25% dos homens e 8% das mulheres admitem assistir a pornografia.

Você também deve ter percebido um pouco desta honestidade nas pesquisas no Google ao ver as formas sugeridas pelo mecanismo de busca quando tenta completar automaticamente o que você digita. Essas sugestões são baseadas nas buscas mais comuns feitas por outras pessoas. Assim, o preenchimento automático nos oferece dicas do que as pessoas pesquisam, mas na verdade ele pode ser um tanto dissimulado. O Google não sugere determinadas palavras consideradas inapropriadas, tais como “pinto”, “foda” e “pornô.” Isso significa que o preenchimento automático nos diz que os pensamentos das pessoas no Google são menos picantes do que de fato são. Mesmo assim, alguns temas delicados aparecem com frequência.

Se você digitar “Por que o…”, as duas primeiras sugestões do Google atualmente são “Por que o céu é azul?” e “Por que existe o dia 29 de fevereiro?”, o que indica que estas são duas das formas mais comuns de completar essa busca. A terceira é: “Por que meu cocô está verde?” E o autopreenchimento do Google pode ficar um tanto estranho. Hoje, se digitar “É normal querer…”, a primeira sugestão é “matar”. Se acrescentar o matar, a primeira sugestão é: “É normal querer matar minha família?”

Precisa de mais provas de que as buscas no Google podem nos dar uma visão diferente do mundo? Pense nas buscas relacionadas a arrependimentos sobre a decisão de ter ou não filhos. Antes de decidir, algumas pessoas temem terem tomado a decisão errada. E, quase sempre, a pergunta é se irão se arrepender de não ter filhos. As pessoas têm sete vezes mais probabilidade de perguntar ao Google se irão se arrepender de não ter do que de ter filhos.

Depois de tomar uma decisão — seja de ter um filho (ou adotar) ou não —, as pessoas às vezes confessam ao Google que lamentam a escolha. Isto pode parecer chocante, mas depois de tomada a decisão, os números são inversos. Adultos com filhos são 3,6 vezes mais propensos a dizer ao Google que se arrependem da decisão que tomaram do que adultos sem filhos.

Uma advertência que é preciso ter em mente ao longo deste capítulo: o Google apresenta uma tendência a expressar pensamentos inadequados, que as pessoas acham que não podem discutir com mais ninguém. Todavia, quando tentamos desvendar pensamentos ocultos, a capacidade do Google de os desmascarar pode ser útil. E a grande disparidade entre ter ou não filhos parece nos mostrar que o pensamento inadequado neste caso é bastante significativo.

Façamos uma pausa para avaliar o que significa fazer uma busca como: “Eu me arrependo de ter tido filhos.” O Google se apresenta como uma fonte em que podemos buscar a informação diretamente, sobre assuntos como o clima, quem venceu o jogo de ontem à noite ou quando a Estátua da Liberdade foi erguida. Mas, às vezes, digitamos

no Google pensamentos sem qualquer censura, sem muita esperança de que ele seja capaz de nos ajudar. Neste caso, a janela de busca serve como um tipo de confessionário.

Há milhares de buscas anualmente, por exemplo, por “Odeio frio”, “Pessoas são insuportáveis” e “Estou triste”. Obviamente, os milhares de buscas por “Estou triste” representam apenas uma minúscula fração das centenas de milhões de pessoas que se sentem tristes em determinado ano. Buscas expressando pensamentos, e não por informações, de acordo com minha pesquisa, são feitas apenas por uma pequena amostra dentre as pessoas que experimentam esses pensamentos. De modo semelhante, minha pesquisa sugere que sete mil buscas que são realizadas por norte-americanos todo ano por “Eu me arrependo de ter tido filhos” representam uma pequena amostra de todas as pessoas que tiveram este pensamento.

Filhos são, obviamente, uma grande alegria para muitas, provavelmente, para a maioria das pessoas. E apesar dos temores de minha mãe de que “você e sua análise de dados idiota” limitassem seu número de netos, esta pesquisa não mudou meu desejo de ter filhos. Mas esse arrependimento inadequado é interessante — e outro aspecto de humanidade que não costumamos ver em conjuntos de dados tradicionais. Nossa cultura constantemente nos inunda com imagens de famílias felizes e maravilhosas. A maioria das pessoas nunca consideraria ter filhos como algo de que pudesse se arrepender. Mas algumas se arrependem. Elas não podem admitir isso para ninguém — exceto para o Google.

A VERDADE SOBRE O SEXO

Quantos homens norte-americanos são gays? Esta é uma questão lendária na pesquisa da sexualidade. Ainda assim tem sido uma das perguntas mais difíceis para os cientistas sociais responderem. Psicólogos não acreditam mais na famosa estimativa feita por Alfred Kinsey — baseada em pesquisas com excessiva amostragem de michês e presidiários — de que 10% dos homens norte-americanos são gays. Pesquisas representativas agora nos dizem que eles são aproximadamente 2% a 3% dos homens. Mas a preferência sexual há muito tempo está entre os assuntos sobre os quais as pessoas tendem a mentir. Acredito que posso usar Big Data para nos dar a melhor resposta que tivemos até hoje para essa pergunta.

Primeiro, vamos falar mais sobre aqueles dados. Pesquisas nos dizem que existem muito mais homens gays em estados tolerantes do que em intolerantes. Por exemplo, de acordo com uma pesquisa do Gallup, a proporção da população que é gay é quase duas vezes maior em Rhode Island, o estado com maior apoio ao casamento gay, do que em Mississippi, o estado com o menor apoio.

Há duas explicações possíveis para isto. Primeiro, homens gays nascidos em estados intolerantes podem se mudar para locais de maior tolerância. Segundo, em estados intolerantes, eles podem não divulgar que são gays; são ainda mais propensos a mentir.

Um pouco de análise sobre a explicação número um — deslocamento dos gays — pode ser colhida em outra fonte de Big Data: o Facebook, que permite aos usuários indicar qual é o gênero de interesse. Cerca de 2,5% dos usuários homens do Facebook que indicam o gênero em que estão interessados dizem que preferem homens, o que corresponde aproximadamente ao que a pesquisa indica. E o Facebook também mostra grandes diferenças na população gay em estados com alta versus baixa tolerância: o Facebook tem uma população gay duas vezes maior em Rhode Island do que em Mississippi.

O Facebook também oferece informações sobre como as pessoas se deslocam no espaço. Fui capaz de codificar a cidade natal de uma amostra de usuários do Facebook assumidamente gays. Isto permitiu estimar diretamente quantos homens gays se mudam de estados intolerantes para partes do país mais tolerantes. A resposta? Claramente há um deslocamento — da cidade de Oklahoma para San Francisco, por exemplo. Mas estimo que os homens encaixotando seus CDs de Judy Garland e partindo para lugares de mentalidade mais aberta explicam menos da metade da diferença na população assumidamente gay em estados tolerantes versus intolerantes.*

Além disso, o Facebook nos permite focar alunos do ensino médio. Este é um grupo especial, pois garotos neste grupo raramente escolhem onde moram. Se o deslocamento explica as diferenças nas populações abertamente homossexuais entre estados, essas diferenças não apareceriam entre usuários do Facebook do ensino médio. Então o que os dados nos dizem? Há muito menos garotos no ensino médio declaradamente gays em estados intolerantes. Apenas dois a cada mil homens deste grupo no Mississippi são assumidamente gays. Portanto, não é apenas o deslocamento.

Se um número semelhante de homens gays nasce em cada estado e o deslocamento não explica totalmente por que alguns estados têm um número maior de homens declaradamente gays, o armário deve desempenhar um papel importante. O que nos leva de volta ao Google, com o qual tantas pessoas mostram disposição para compartilhar tanto de suas vidas.

Existe uma maneira de usar as buscas por pornografia para testar quantos homens gays realmente existem em diferentes estados? De fato, existe. Em todo o país, estimo — usando os dados de busca no Google e no Google AdWords — que cerca de 5% das buscas de homens por pornografia são por pornografia masculina gay. (Este número incluiria as buscas por termos como "Rocket Tube", um site pornô gay popular, assim como por "gay porno".)

E como isso varia em diferentes partes do país? De modo geral, há mais buscas por pornografia gay em estados tolerantes em comparação aos intolerantes. Isto faz sentido, considerando-se que alguns homens gays se mudam de locais intolerantes para outros tolerantes. Mas as diferenças não são nem de perto tão grandes quanto as sugeridas pela pesquisa ou pelo Facebook. No Mississippi, estimo que 4,8% das buscas feitas por homens por pornografia são por pornô gay, muito superiores aos números sugeridos pela pesquisa ou pelo Facebook, e razoavelmente próximos aos 5,2% das buscas por pornografia gay em Rhode Island.

Então, quantos homens norte-americanos são gays? Esta medição das buscas por pornografia realizadas por homens — aproximadamente 5% são pelo mesmo sexo — parece uma estimativa razoável do tamanho real da população gay nos Estados Unidos. E há outra maneira, menos direta, de obter esse número. Ela requer um pouco de ciência de dados. Podemos utilizar a relação entre tolerância e a população declaradamente gay. Acompanhe meu raciocínio.

Minha pesquisa preliminar indica que em determinado estado cada vinte pontos percentuais de apoio ao casamento gay significa cerca de uma vez e meia o número de homens naquele estado que se identificarão abertamente como gays no Facebook. Com base nisto, podemos estimar quantos homens nascidos em uma região hipoteticamente tolerante — onde, digamos 100% das pessoas apoiem o casamento gay — seriam assumidamente gays. Minha estimativa é que esse número seria de 5%, o que se encaixa perfeitamente nos dados das buscas por pornografia. O ambiente mais tolerante em que se pode crescer é o dos garotos do ensino médio da região da baía da Califórnia. Cerca de 4% deles são abertamente gays no Facebook. Isto parece se alinhar com meus cálculos.

Devo observar que não fui capaz de fazer uma estimativa sobre a homossexualidade feminina. Os números sobre pornografia são menos úteis neste caso, pois muito menos mulheres assistem a pornografia, tornando a amostra menos representativa. E entre as que assistem, mesmo as que são a princípio atraídas por homens na vida real parecem gostar de assistir a pornografia lésbica. Vinte por cento do total de vídeos assistidos por mulheres no PornHub são lésbicos.

Cinco por cento dos homens norte-americanos serem gays é uma estimativa, é claro. Alguns são bissexuais; alguns — especialmente quando jovens — ainda não têm certeza sobre sua sexualidade. Obviamente, não podemos considerá-la tão precisa quanto o número de pessoas que votaram ou que assistiram a um filme.

Mas uma consequência de minha estimativa é clara: um grupo impressionante de homens nos Estados Unidos, especialmente em estados intolerantes, ainda está no armário. Eles não revelam suas preferências sexuais no Facebook. Não as divulgam em pesquisas. E em muitos casos, até se casam com mulheres.

Ocorre que as esposas desconfiam que seus maridos são gays com certa frequência. Elas demonstram esta suspeita na busca surpreendentemente comum: "Meu marido é gay?" "Gay" é 10% mais frequente como complemento de buscas começando por "Meu marido é/está…" do que a segunda opção, "me traindo". É oito vezes mais comum do que "[é] alcoólatra" e dez do que "[está] deprimido".

O mais revelador, talvez, seja o fato de as buscas questionando a sexualidade dos maridos serem muito mais prevalentes em regiões menos tolerantes. Os estados com o maior porcentual de mulheres fazendo essa pergunta são a Carolina do Sul e a Louisiana. Na verdade, em 21 dos 25 estados em que essa pergunta é mais comum, o apoio ao casamento gay é mais baixo do que a média nacional.

O Google e os sites pornôs não são os únicos recursos de dados quando se trata da sexualidade de homens. Há mais evidências disponíveis em Big Data sobre o que significa viver no armário. Analisei anúncios no Craigslist de homens buscando por "encontros casuais". O porcentual desses anúncios para encontros casuais com outros homens tende a ser superior em estados menos intolerantes. Entre os estados com os maiores porcentuais estão Kentucky, Louisiana e Alabama.

E para uma visão melhor do interior do armário, vamos retornar aos dados de busca no Google e ser um pouco mais específicos. Uma das expressões mais comuns feitas imediatamente antes ou depois de "pornô gay" é "teste gay". (Esses testes alegam indicar se um homem é ou não homossexual.) E as buscas por "teste gay" são duas vezes mais prevalentes nos estados menos tolerantes.

O que significa fazer as duas buscas, por "pornô gay" e "teste gay", em sequência? Aparentemente, isto sugere uma mente bastante confusa ou até torturada. É razoável suspeitar que esses homens esperam confirmar que seu interesse por pornografia gay não significa necessariamente que são homossexuais.

Os dados de busca do Google não nos permitem ver o histórico de busca de um usuário específico. Entretanto, em 2006, a AOL liberou uma amostra das buscas de seus usuários para pesquisadores acadêmicos. Veja a seguir as

buscas de um usuário anônimo em um período de seis dias.

Sexta-feira 03h49m55	fotos gays gratis [sic]
Sexta-feira 03h59m37	fotos gay vestiário
Sexta-feira 04h00m14	fotos gay
Sexta-feira 04h00m35	fotos sexo gay
Sexta-feira 05h08m23	teste gay longo
Sexta-feira 05h10m00	bom teste gay
Sexta-feira 05h25m07	testes gay para homens confusos
Sexta-feira 05h26m38	testes gay
Sexta-feira 05h27m22	testes eu sou gay
Sexta-feira 05h29m18	fotos gay
Sexta-feira 05h30m01	fotos homens nus
Sexta-feira 05h32m27	fotos homens nus grátis
Sexta-feira 05h38m19	fotos gay sexo
Sexta-feira 05h41m34	sexo fotos bundas homens gostosos
Quarta-feira 13h37m37	testes eu sou gay
Quarta-feira 13h41m20	gay testes
Quarta-feira 13h47m49	homens bundas sexo
Quarta-feira 13h50m31	videos gratis sexo gay [sic]

Certamente, os dados mostram um homem que não está confortável com sua sexualidade. E os dados do Google nos dizem que ainda existem muitos homens como ele. A maior parte deles, de fato, vive em estados menos tolerantes a relacionamentos entre pessoas do mesmo sexo.

Para um olhar mais detalhado nas pessoas por trás desses números, perguntei a um psiquiatra do Mississippi, especialista em ajudar homens confusos sobre a sexualidade, se algum de seus pacientes aceitaria conversar comigo. Um deles concordou. Ele me disse que era um professor aposentado, na faixa dos sessenta anos, casado com a mesma mulher há mais de quarenta.

Cerca de dez anos antes, sobrecarregado por causa do estresse, consultou o psiquiatra e finalmente reconheceu sua sexualidade. Ele conta que sempre soube que sentia atração por homens, mas pensava que isso fosse universal e algo que todo homem na verdade escondia. Logo depois de começar a terapia, ele teve seu primeiro, e único, encontro homossexual, com um aluno de pouco menos de trinta anos, uma experiência que descreve como “maravilhosa”.

Ele e a mulher não fazem sexo. Ele diz que se sentiria culpado se terminasse seu casamento ou saísse com homens abertamente. Ele se arrepende de praticamente todas as decisões importantes de sua vida.

O professor aposentado e sua mulher continuam a viver noites sem amor romântico, sem sexo. Apesar dos enormes progressos, a persistência da intolerância fará com que milhões de outros norte-americanos façam o mesmo.

Você não deve ter se chocado com a informação de que 5% dos homens são gays e que muitos ainda continuam no armário. Já houve tempo em que a maioria das pessoas ficaria chocada. E ainda existem lugares em que muitas ainda ficam.

“No Irã não temos homossexuais como em seu país”, declarou Mahmoud Ahmadinejad, então presidente do Irã, em 2007. “No Irã não temos este fenômeno.” Do mesmo modo, Anatoly Pakhomov, prefeito de Sochi, Rússia, pouco antes de a cidade ser sede dos Jogos Olímpicos de Inverno em 2014, disse: “Não temos gays em nossa cidade.” Ainda assim o comportamento na internet revela significativo interesse por pornografia gay em Sochi e no Irã.

Isso levanta uma pergunta óbvia: existe algum tipo de interesse sexual comum nos Estados Unidos que ainda é considerado chocante hoje? Depende do que você considera comum e com que facilidade é capaz de se chocar.

Grande parte das buscas mais comuns no PornHub são bastante óbvias — elas incluem termos como “teen” [adolescentes], “threesome” [sexo a três] e “blowjob” [boquete] para homens, frases como “passionate love making” [fazer amor apaixonadamente], “nipple sucking” [chupando mamilos] e “man eating pussy” [homem chupando xoxota] para mulheres.

Fora das buscas tradicionais, os dados do PornHub nos mostram alguns fetiches que você não adivinharia que existem. Há mulheres fazendo buscas por “anal apples” [maçãs anais] e “humping stuffed animals” [transando com animais de pelúcia]. Há homens buscando por “snot fetish” [fetiche por meleca] e “nude crucifixion” [crucificação nua]. Mas estas buscas são raras — apenas cerca de dez por mês mesmo em um site grande como esse.

Outra questão relacionada que se torna muito clara ao analisar os dados do PornHub: há pessoas para todos os gostos. Mulheres, o que não é de surpreender, buscam com frequência por termos como homens “altos”, “morenos” e “lindos”. Mas elas também às vezes buscam por homens “baixos”, “branquelos” e “feios”. Há mulheres que buscam por homens “mutilados”, “homens gordos com pinto pequeno” e “homens gordos, velhos e feios”. Homens frequentemente buscam por mulheres “magras”, mulheres com “peitos grandes” e mulheres “loiras”. Mas eles também às vezes buscam por mulheres “gordas”, com “peitos bem pequenos” e mulheres com “cabelo verde”. Há homens que buscam por mulheres “carecas”, “anãs” e mulheres “sem mamilos”. Estes dados podem ser animadores para aqueles que não são altos, morenos e lindos, ou aquelas que não são magras, louras e com seios fartos*.

E quanto às buscas que são comuns e surpreendentes ao mesmo tempo? Entre as 150 buscas mais comuns feitas por homens, as mais inusitadas são as incestuosas, discutidas no Capítulo 2, que fala sobre Freud. Outros focos dos desejos masculinos pouco discutidos são “travestis” (77ª busca mais comum) e “vovós” (110ª). Em geral, cerca de 1,4% das buscas feitas por homens no PornHub é por mulheres com pênis. Cerca de 0,6% (0,4% dos homens abaixo dos 34 anos) é por idosas. Apenas 1 em 24 mil buscas do PornHub feitas por homens são explicitamente por pré-adolescentes; o que pode estar relacionado ao fato de que, por motivos óbvios, o PornHub, bane todas as formas de pornografia infantil e informa que é ilegal.

Dentre as buscas mais populares no PornHub feitas por mulheres está um gênero de pornografia que, aviso de antemão, irá perturbar muitos leitores: imagens de sexo com violência contra mulheres. Pelo menos 25% das mulheres buscam por pornografia heterossexual que enfatize a dor ou a humilhação de uma mulher — “anal doloroso chorando”, “humilhação pública” e “sexo grupal extremo violento” por exemplo. Cinco por cento procuram por sexo não consensual — “estupro” ou sexo “forçado” — mesmo esses vídeos sendo banidos pelo PornHub. E as taxas de busca por todos estes termos são pelo menos duas vezes mais comuns entre mulheres do que entre homens. Se existe um gênero de pornografia em que a violência é perpetrada contra uma mulher, minha análise dos dados mostra que ele é quase sempre desproporcionalmente mais atraente para mulheres.

Obviamente, ao tentar compreender isso, é realmente importante lembrar que existe uma grande diferença entre fantasias e vida real. Sim, dentre a minoria de mulheres que visitam o PornHub, há um subconjunto que busca — sem sucesso — por imagens de estupro. Para dizer o óbvio, isto não significa que as mulheres queiram ser estupradas na vida real, e certamente não torna o estupro um crime menos hediondo. O que os dados pornográficos nos dizem é que, às vezes, as pessoas têm fantasias que gostariam de não ter, e que podem nunca admiti-las para alguém. Armários não são apenas depósitos de fantasias. Quando se trata de sexo, as pessoas guardam muitos segredos — como a frequência com que fazem, por exemplo.

Na introdução, comentei que os norte-americanos relatam usar muito mais preservativos do que a quantidade vendida por ano. Assim, você pode pensar que isto significa que as pessoas dizem usar preservativos com mais frequência do que realmente usam. As evidências sugerem que elas exageram na quantidade de vezes que fazem sexo, também. Aproximadamente 11% das mulheres entre 15 e 44 anos declaram ser sexualmente ativas, não estar grávidas e não usar contraceptivo. Mesmo com suposições relativamente conservadoras sobre quantas vezes fazem sexo, cientistas esperariam que 10% delas ficassem grávidas todo mês. Mas este já seria um número maior do que o total de gravidezes nos Estados Unidos (que é de 1 para cada 113 mulheres em idade fértil). Em nossa cultura obcecada por sexo pode ser difícil admitir que não se está transando muito.

Mas se estiver buscando por compreensão ou conselho, você tem, mais uma vez, um incentivo para pedir ao Google. Lá, há dezesseis vezes mais reclamações sobre um(a) parceiro(a) casado(a) que não quer fazer sexo do que de um que não está disposto a conversar. Existem cinco vezes e meia mais reclamações sobre um parceiro não casado que não quer fazer sexo do que um que não responda mensagens.

E as buscas no Google sugerem um inesperado culpado para muitos desses relacionamentos sem sexo. Há duas vezes mais reclamações de namorados do que de namoradas que não querem transar. De longe, a busca mais comum sobre um namorado é: “Meu namorado não quer transar comigo.” (As buscas no Google não são divididas por gênero, mas, tendo em vista que a análise anterior mostra que 95% dos homens são héteros, podemos supor que muitas buscas por “namorado” não são feitas por homens.)

Como devemos interpretar isto? Isto realmente indica que namorados evitam sexo mais do que namoradas? Não necessariamente. Como mencionei anteriormente, as buscas no Google são tendenciosas para assuntos que as pessoas se sentem mais incomodadas em falar. Homens podem se sentir mais confortáveis em dizer aos amigos sobre a falta de interesse sexual de suas namoradas do que mulheres. Ainda assim, mesmo que os dados do Google não indiquem que os namorados tenham duas vezes mais propensão a evitar sexo com suas namoradas, eles sugerem que namorados que não querem transar são mais comuns do que as pessoas assumem.

Os dados do Google ainda sugerem um dos motivos para as pessoas estarem evitando o sexo com tanta frequência: uma enorme ansiedade, muita dela injustificada. Vamos começar com a ansiedade dos homens. Não é novidade que homens se preocupam com o quanto são bem-do-tados, mas o grau desta preocupação é bastante intenso.

Homens pesquisam no Google mais perguntas sobre seus órgãos sexuais do que sobre qualquer outra parte do corpo: mais que seus pulmões, fígado, pés, orelhas, nariz, garganta e cérebro combinados. Homens fazem mais buscas por como aumentar o tamanho do pênis do que sobre como afinar o violão ou trocar um pneu. A principal preocupação pesquisada sobre o uso de esteroides não é se ele prejudica a saúde, mas sim se pode diminuir o tamanho do pênis. A pergunta mais buscada no Google relacionada a como o corpo ou a mente deles mudará conforme envelhecem foi se o tamanho do pênis diminui.

Nota: Uma das perguntas mais comuns no Google em relação à genitália masculina é: “O meu pênis é grande?” Esses homens recorrem ao Google, em vez de usar uma régua, para perguntar, isso é, em minha opinião, uma expressão típica de nossa era digital.*

As mulheres se preocupam com o tamanho do pênis? Raramente, de acordo com as buscas no Google. Para cada busca que as mulheres fazem sobre o membro do parceiro, aproximadamente 170 buscas são feitas por homens. É verdade que, nas raras ocasiões em que as mulheres expressam sua preocupação quanto ao pênis dos parceiros, frequentemente é sobre o tamanho, mas não necessariamente por ser pequeno. Mais de 40% das reclamações sobre o tamanho do pênis dos parceiros é por ser grande demais. “Dor” é a palavra mais usada em buscas com a frase “durante o sexo”. (“Sangrar”, “fazer xixi”, “chorar” e “soltar pum” são as outras quatro mais comuns.) Ainda assim 1% das buscas feitas por homens querendo mudar o tamanho do pênis busca informações sobre como o diminuir.

A segunda pergunta sexual mais comum entre homens é como fazer o ato sexual durar mais. De novo, as inseguranças dos homens não parecem combinar com as preocupações das mulheres. Há praticamente o mesmo número de buscas perguntando como fazer o namorado chegar ao clímax mais rápido e mais devagar. Na verdade, a preocupação mais comum das mulheres em relação ao orgasmo dos namorados não é sobre quando ocorre, mas por que não acontece.

Homens não costumam falar sobre questões de problemas de imagem corporal. E enquanto é verdade que o interesse geral na aparência física tende a ser feminino, não é tão desequilibrado como o estereótipo sugere. De acordo com minha análise do Google AdWords, que mede os sites que as pessoas visitam, o interesse por beleza e fitness entre os homens é de 42%; perda de peso, 33% e cirurgias estéticas, 39%. Dentre todas as buscas com “como” relacionadas a seios, cerca de 20% quer saber como se livrar da ginecomastia.

Mas mesmo que o número de homens que carecem de confiança em seus corpos seja maior do que a maioria das pessoas pensa, as mulheres ainda os superam no que diz respeito a inseguranças sobre a aparência. Então, o que esse soro da verdade digital revela sobre as inseguranças femininas? Todo ano, nos Estados Unidos, há mais de sete milhões de buscas procurando por próteses nos seios. A estatística oficial nos mostra que cerca de 300 mil mulheres se submetem a esse procedimento anualmente.

Mulheres também mostram uma significativa insegurança sobre as nádegas, embora muitas tenham mudado de opinião sobre o que não gostam no próprio corpo.

Em 2004, em algumas partes dos Estados Unidos, a busca mais comum em relação a mudanças no bumbum era sobre como diminuí-lo. O desejo de aumentar as nádegas era amplamente concentrado em áreas com grande população negra. No início de 2010, porém, o desejo por bumbuns maiores cresceu nos Estados Unidos. Este interesse, e o próprio tamanho dos glúteos, triplicou em quatro anos. Em 2014, havia mais buscas sobre como aumentar o bumbum do que como diminuir em todos os estados. Atualmente, para cada cinco buscas procurando por próteses de seios nos Estados Unidos, uma é para próteses de bumbum. (Obrigado, Kim Kardashian!)

O aumento da preferência das mulheres por bumbuns maiores equivale à dos homens? Curiosamente, sim. Buscas por “pornô bunda grande”, que também costumavam se concentrar em comunidades negras, recentemente,

dispararam em popularidade nos Estados Unidos.

O que mais os homens querem no corpo de uma mulher? Como mencionei anteriormente, e como muitos acharão incrivelmente óbvio, homens mostram uma preferência por seios grandes. Cerca de 12% das buscas pornográficas não genéricas é por seios fartos. Isso são vinte vezes mais do que o volume de buscas pornográficas por seios pequenos.

Não está claro que isto signifique que homens queiram que as mulheres coloquem próteses nos seios. Cerca de 3% das buscas pornôs por seios grandes declaram explicitamente querer seios naturais.

As buscas no Google envolvendo a esposa e próteses mamárias são divididas igualmente entre perguntas de como a persuadir a colocar e a perplexidade de por que elas querem colocá-las.

Ou pense na busca mais comum sobre os seios da namorada: “Eu adoro os seios da minha namorada.” Não está claro o que os homens esperam encontrar no Google ao fazer esta busca.

Mulheres, assim como homens, têm dúvidas sobre seus órgãos genitais. Na verdade, elas têm o mesmo número de perguntas sobre suas vaginas que os homens sobre seus pênis. As preocupações das mulheres sobre suas vaginas frequentemente são relacionadas à saúde. Mas pelo menos 30% das perguntas envolvem outras preocupações. Mulheres querem saber como depilar, estreitar e melhorar o gosto de suas vaginas. Uma preocupação notavelmente comum, já mencionada brevemente, é sobre como melhorar o odor.

Mulheres se preocupam com mais frequência que suas vaginas tenham cheiro de peixe, seguido por vinagre, cebola, amônia, alho, queijo, cheiro de corpo, urina, pão, água sanitária, fezes, suor, metal, chulé, lixo e carne estragada.

Em geral, homens não fazem muitas buscas no Google envolvendo a genitália da parceira. Homens fazem aproximadamente o mesmo número de buscas sobre a vagina das namoradas do que as mulheres sobre os pênis do namorado.

Quando homens buscam sobre a vagina da parceira, é normalmente para reclamar sobre o que elas mais se preocupam: o cheiro. Na maior parte, os homens estão tentando descobrir como dizer à mulher sobre o mau cheiro sem ferir seus sentimentos. Às vezes, porém, as perguntas dos homens sobre o odor revelam suas próprias inseguranças. Homens ocasionalmente perguntam por maneiras de usar o cheiro para detectar traição — se cheira a preservativos, por exemplo, ou ao sêmen de outro homem.

O que podemos extrair de toda essa insegurança oculta? Há claramente boas notícias aqui. O Google nos oferece razões legítimas para nos preocupar menos. Muitos de nossos medos mais profundos sobre como nossos parceiros sexuais nos percebem não são justificados. Sozinhos, em seus computadores, sem incentivo para mentir, parceiros se revelam bastante indulgentes e não superficiais. Na verdade, estamos sempre tão ocupados julgando nossos próprios corpos que resta pouca energia para julgar os outros.

Há também provavelmente uma conexão entre duas das principais preocupações reveladas nas buscas sexuais no Google: falta de sexo e uma insegurança sobre a atratividade e o desempenho sexual. Talvez essas estejam relacionadas. Talvez se nos preocupássemos menos com sexo, faríamos mais.

O que mais as buscas no Google nos dizem sobre sexo? Podemos fazer uma batalha dos sexos, para ver quem é mais generoso. Pegue todas as buscas procurando por formas de melhorar a performance no sexo oral no sexo oposto. Homens ou mulheres procuram por mais dicas? Quem é sexualmente mais generoso, homens ou mulheres? Mulheres, dãã. Somando-se todas as possibilidades, estimo que a proporção é de 2:1 em favor das mulheres buscando por conselhos sobre como fazer sexo oral melhor em seus parceiros.

E quando os homens procuram por dicas de como fazer sexo oral, frequentemente não estão procurando por maneiras de agradar outra pessoa. Homens fazem o mesmo número de buscas procurando por maneiras de realizar sexo oral em si mesmos quanto de como fazer uma mulher atingir o orgasmo. (Este está entre meus fatos favoritos dos dados de busca no Google.)

A VERDADE SOBRE ÓDIO E PRECONCEITO

Sexo e romance são dificilmente os únicos assuntos escondidos pela vergonha e, portanto, não são exclusivos sobre as pessoas guardarem segredo. Muitas pessoas são, por uma grande razão, inclinadas a guardar seus preconceitos para si mesmas. Suponho que se possa chamar de progresso o fato de muitas pessoas hoje sentirem que serão julgadas se admitirem que julgam outras pessoas com base em sua etnia, orientação sexual ou religião. Mas muitos norte-americanos ainda fazem isso. (Esta é outra seção, aviso aos leitores, que inclui material um tanto perturbador.)

Você pode ver isso no Google, onde os usuários às vezes perguntam coisas como: “Por que pessoas negras são grosseiras”, ou “Por que judeus são perversos?” Abaixo, em ordem, estão as cinco principais palavras negativas usadas em buscas sobre vários grupos.

	1.	2.	3.	4.	5.
AFRO-AMERICANOS	grosseiro	racista	burro	feio	preguiçoso
JUDEUS	perverso	racista	feio	mesquinho	ganancioso
MUÇULMANOS	perverso	terrorista	mau	violento	perigoso
MEXICANOS	racista	burro	feio	preguiçoso	idiota
ASIÁTICOS	feio	racista	irritante	burro	mesquinho
GAYS	perverso	errado	burro	irritante	egoísta
CRISTÃOS	burro	louco	idiota	lunático	errado

Alguns poucos padrões entre estes estereótipos se destacam. Por exemplo, afro-americanos são o único grupo que sofre com o estereótipo de "grosseiro". Praticamente todo grupo é vítima do estereótipo "burro"; os únicos dois que não são vistos assim são judeus e muçulmanos. O estereótipo "perverso" é empregado para judeus, muçulmanos e gays, mas não para negros, mexicanos, asiáticos e cristãos.

Muçulmanos são o único grupo estereotipado como terrorista. Quando um muçulmano norte-americano age de acordo com este estereótipo, a resposta pode ser instantânea e cruel. Os dados de busca do Google nos dão uma visão minuto a minuto nestas erupções de raiva impulsionadas pelo ódio.

Analisemos o que aconteceu logo depois do tiroteio em massa em San Bernardino, Califórnia, em 2 de dezembro de 2015. Naquela manhã, Rizwan Farook e Tashfeen Malik entraram em uma reunião de colegas de trabalho de Farook armados com pistolas e rifles semiautomáticos e mataram catorze pessoas. Naquela noite, literalmente minutos após o primeiro relato da mídia sobre o nome muçulmano de um dos atiradores, um perturbador número de californianos já havia decidido o que queria fazer com os muçulmanos: matar.

Na Califórnia, a principal busca no Google contendo a palavra "muçulmanos" naquele momento era "matar muçulmanos". E em geral, norte-americanos buscaram pela frase "matar muçulmanos" com aproximadamente a mesma frequência que buscaram por "receita de martini", "sintomas de enxaqueca" e "escalação dos Cowboys". Nos dias seguintes ao ataque em San Bernardino, para cada norte-americano preocupado com a "islamofobia", outro buscava por "matar muçulmanos". Enquanto as buscas de ódio eram aproximadamente 20% de todas as buscas sobre muçulmanos antes do ataque, mais da metade de todo o volume de busca sobre muçulmanos se tornou de ódios nas horas que seguiram ao ataque.

E estes dados de busca minuto a minuto nos dizem o quão difícil pode ser aplacar este ódio. Quatro dias depois do tiroteio, o então presidente Obama fez um comunicado à nação em horário nobre. Ele queria reassegurar aos norte-americanos que o governo poderia acabar com o terrorismo e, talvez o mais importante, acalmar esta perigosa islamofobia.

Obama apelou aos nossos melhores instintos, falando da importância da inclusão e da tolerância. A retórica foi poderosa e comovente. O *Los Angeles Times* elogiou Obama por: "[nos alertar] contra permitir que nossos medos obscureçam nosso julgamento." O *New York Times* chamou o discurso de "rígido" e "apaziguador". O site "Think Progress" elogiou-o como: "Uma ferramenta necessária para boa governança, direcionada para salvar vidas de muçulmanos norte-americanos." O discurso de Obama, em outras palavras, foi considerado um grande sucesso. Mas foi de fato?

Os dados de busca do Google sugerem o contrário. Junto com Evan Soltas, então em Princeton, examinei os dados. Em seu discurso, o presidente disse: "É responsabilidade de todos os norte-americanos — de todas as crenças — rejeitar a discriminação." Mas as buscas chamando muçulmanos de "terroristas", "maus", "violentos" e "perversos" duplicaram durante e logo depois do discurso. O presidente Obama também disse: "É nossa responsabilidade rejeitar testes religiosos sobre quem aceitamos em nosso país." Mas buscas negativas sobre refugiados sírios, um grupo composto em grande parte por muçulmanos que então procurava desesperadamente por um refúgio seguro, aumentaram em 60%, enquanto as buscas perguntando como os ajudar diminuíram 35%. Obama pediu aos norte-americanos para "não esquecerem que a liberdade é mais poderosa que o medo". Ainda assim, buscas por "matar muçulmanos" triplicaram durante seu discurso. Na verdade, praticamente toda palavra negativa que pudemos pensar em testar em relação a muçulmanos surgiu durante ou depois do discurso de Obama, e todas as buscas positivas que pudemos pensar em testar diminuíram.

Em outras palavras, Obama pareceu dizer todas as coisas certas. Toda a mídia tradicional parabenizou Obama por suas palavras curativas. Mas os novos dados da internet, oferecendo o soro da verdade digital, sugeriram que o discurso na verdade teve o efeito oposto. Em vez de acalmar a multidão enfurecida, como todos pensaram que ele estava fazendo, os dados da internet nos mostram que Obama na verdade inflamou os ânimos. Coisas que pensamos que funcionam podem ter o efeito exatamente oposto do esperado. Às vezes, precisamos dos dados da internet para corrigir nosso instinto de nos congratular.

Então, o que Obama deveria ter dito para aplacar esta forma específica de ódio atualmente tão nociva nos Estados Unidos? Voltaremos a isto mais adiante. Agora vamos analisar esse antigo veio de preconceito nos Estados Unidos, a forma do ódio que na verdade se destaca de todo o resto, aquela que tem sido a mais destrutiva e assunto da pesquisa que originou este livro. Em meu trabalho com os dados de busca do Google, o fato mais revelador que descobri em relação ao ódio na internet é a popularidade da palavra "nigger".

Seja na forma singular ou plural, a palavra "nigger" é incluída em sete milhões de buscas nos Estados Unidos por ano. (Repito, a palavra usada em letras de rap é quase sempre "nigga", não "nigger", então as letras de hip-hop não têm impacto significativo na contabilização.) As buscas por "nigger jokes" [piadas de crioulo] são dezessete vezes mais comuns do que as buscas por piadas com judeus, asiáticos, latinos, chineses e gays, usando termos pejorativos, combinadas.

Quando as buscas por "nigger(s)" — ou "nigger jokes" [piadas de crioulos] — são mais comuns? Sempre que um afro-americano é notícia. Entre os períodos em que essas buscas foram mais altas está o momento logo após o Furacão Katrina, quando a televisão e os jornais mostravam imagens de pessoas negras desesperadas em Nova Orleans lutando para sobreviver. Elas também dispararam durante a primeira eleição de Obama. E as buscas por "nigger jokes" [piadas de crioulos] aumentam em média 30% no Dia de Martin Luther King Jr.

A assustadora ubiquidade dessa difamação racial coloca em dúvida um pouco do entendimento atual do racismo.

Qualquer teoria de racismo tem que explicar um grande quebra-cabeça nos Estados Unidos. Por outro lado, a esmagadora maioria de afro-americanos acha que sofre preconceito — e eles têm ampla evidência da discriminação em revistas policiais, entrevistas de emprego e decisões de júri. Por outro lado, raros são os norte-americanos brancos que admitem ser racistas.

A explicação dominante entre os cientistas políticos atualmente é de que isso se deve, em grande parte, à ampla disseminação do preconceito *implícito*. Norte-americanos brancos podem não ter a intenção, de acordo com esta teoria, mas têm um preconceito subconsciente, que influencia o modo como tratam os afro-americanos. Acadêmicos inventaram uma forma engenhosa de testar este preconceito. Ele é chamado de teste de associação-implícita.

Os testes mostram de modo consistente que a maioria das pessoas demora milissegundos para associar rostos de pessoas negras com palavras positivas, como "bom", do que com palavras negativas, como "terrível". Para rostos de pessoas brancas, o padrão é inverso. O tempo adicional é evidência do preconceito implícito da pessoa — um preconceito de que a pessoa pode nem ao menos ter consciência.

Existe, porém, uma explicação alternativa para a discriminação que afro-americanos sentem e que brancos negam: o racismo *explícito* velado. Suponha que exista um racismo consciente razoavelmente disseminado do qual as pessoas estão bastante cientes, mas que não confessam — certamente não em uma pesquisa. É isso que os dados de busca parecem dizer. Não há nada de implícito sobre buscar "nigger jokes" [piadas de crioulos]. E é difícil

imaginar que norte-americanos pesquisando no Google a palavra “nigger” com a mesma frequência que “enxaqueca” e “economista” sem racismo *explícito* não tenha um grande impacto sobre os norte-americanos. Antes dos dados do Google, não tínhamos uma medida contundente deste ódio pernicioso. Agora a temos. Estamos, portanto, em posição de ver o que isto explica.

Como discutimos anteriormente, isso explica por que o total de votos de Obama em 2008 e 2012 foi reduzido em muitas regiões. Isso também se correlaciona com a diferença de salário entre negros e brancos, como relatou recentemente uma equipe de economistas. As áreas que descobri fazerem a maior parte das buscas racistas, em outras palavras, pagam menos a pessoas negras. E então, existe o fenômeno da candidatura de Donald Trump. Como observado na introdução, quando Nate Silver, o guru das pesquisas eleitorais, procurou pela variável geográfica com a correlação mais forte com o apoio a Trump nas prévias do Partido Republicano de 2016, encontrei no mapa do racismo que desenvolvi. Esta variável foi buscar por “nigger(s)”.

Acadêmicos recentemente coletaram uma medição estado por estado do preconceito implícito contra negros, o que me permitiu comparar os efeitos do racismo explícito, conforme medido pelas buscas no Google, e o preconceito *implícito*. Por exemplo, testei o quanto cada um trabalhou contra Obama em ambas as eleições presidenciais. Usando a análise regressiva, descobri que as buscas racistas no Google em uma área serviram muito para prever onde Obama teria um baixo desempenho. O desempenho em uma área nos testes de associação-implícita acrescentou pouco.

Para provocar e encorajar mais pesquisas na área, permita-me apresentar a seguinte conjectura, pronta para ser testada pelos acadêmicos em uma ampla gama de campos. A explicação primária para a discriminação contra afro-americanos hoje não é o fato de que as pessoas que concordam em participar de experimentos de laboratório fazem associações subconscientes entre palavras negativas e pessoas negras, mas o fato de milhões de norte-americanos brancos continuarem a fazer coisas como buscar por “nigger jokes” [piadas de crioulos].

A discriminação que as pessoas negras experimentam regularmente nos Estados Unidos parece ser incentivada mais amplamente por hostilidade explícita, mas dissimulada. Mas, para outros grupos, o preconceito subconsciente pode ter um impacto mais fundamental. Por exemplo, fui capaz de usar as buscas no Google para encontrar evidências de preconceito implícito contra outro segmento da população: meninas.

E quem, você pode se perguntar, esconderia preconceito contra meninas?

Os pais delas.

Não é de surpreender que pais costumem se empolgar com a ideia de seus filhos serem brilhantes. Na verdade, de todas as buscas no Google começando por “Meu bebê de 2 anos”, a palavra mais comum para completar é “brilhante”. Mas esta pergunta não é feita igualmente sobre meninos e meninas. Os pais são duas vezes e meia mais propensos a perguntar: “Meu filho é brilhante?” do que “Minha filha é brilhante?” Os pais demonstram um preconceito parecido ao usar outras frases relacionadas à inteligência que podem evitar dizer em voz alta, como: “Meu filho é um gênio?”

Os pais identificam diferenças legítimas entre meninas e meninos? Talvez meninos tenham maior probabilidade do que meninas de usar palavras mais complexas, ou mostram sinais objetivos de genialidade de outra forma? Não. Se existe alguma diferença, é o oposto. Na primeira infância, as meninas apresentam consistentemente vocabulário maior e usam sentenças mais complexas. Em colégios dos Estados Unidos, meninas têm 9% mais probabilidade do que os meninos de entrar em programas para crianças superdotadas. Apesar de tudo isto, pais procuram na mesa de jantar mais por meninos do que meninas brilhantes.* Na verdade, em cada busca de termos relacionados à inteligência que testei, incluindo aquelas indicando ausência, os pais tinham maior probabilidade de perguntar sobre seus filhos do que sobre suas filhas. Há ainda mais buscas por “meu filho está atrasado” ou “é burro” do que as pesquisas equivalentes sobre filhas. Mas as buscas com palavras negativas como “atrasado” e “burro” são menos inclinadas especificamente aos filhos do que por palavras positivas, como “brilhante” ou “gênio”.

Quais são então as preocupações predominantes dos pais em relação a suas filhas? Principalmente, qualquer coisa relacionada à aparência. Considere questões sobre o peso de uma criança. Pais pesquisam no Google: “Minha filha está com excesso de peso?” aproximadamente duas vezes mais do que “Meu filho está com sobrepeso?” Pais têm duas vezes mais probabilidade de perguntar como fazer com que as filhas percam peso do que os filhos. Assim como em relação à genialidade, esta tendência de gênero não é baseada na realidade. Cerca de 28% das meninas e 35% dos meninos estão acima do peso. Mesmo as balanças mostrando mais garotos do que garotas com sobrepeso, os pais veem — ou se preocupam — com o sobrepeso das meninas com muito mais frequência do que com o dos meninos.

Os pais também têm uma vez e meia mais probabilidade de perguntar se suas filhas são bonitas do que se seus filhos o são. E têm quase três vezes mais probabilidade de perguntar se suas filhas são feias do que seus filhos. (Como se espera que o Google saiba se uma criança é bonita ou feia é difícil dizer.)

Em geral, os pais parecem mais propensos a usar palavras positivas em perguntas sobre os filhos. São mais inclinados a perguntar se um filho é "feliz" e menos a perguntar se está "deprimido".

O leitor liberal pode imaginar que essas inclinações são mais comuns em partes conservadoras do país, mas não encontrei provas disto. Na verdade, não achei uma relação significativa entre qualquer dessas tendências e a composição política ou cultural de um estado. Nem existem provas de que essas tendências diminuíram desde 2004, o ano em que os dados de busca do Google foram disponibilizados pela primeira vez. Parece que esse preconceito contra as meninas é mais disseminado e profundamente arraigado do que pensamos.

O sexismo não é a única coisa sobre a qual nossos estereótipos a respeito do preconceito podem estar equivocados.

Vikingmaiden88 tem 26 anos de idade. Ela gosta de ler sobre história e escrever poesia. Sua citação na assinatura é de Shakespeare. Eu colhi isso de seu perfil e posts do Stormfront.org, o site de ódio mais popular dos Estados Unidos. Também descobri que Vikingmaiden88 gostou do conteúdo do site do jornal em que trabalho, o *New York Times*. Ela escreveu um post entusiasmado sobre uma coluna específica do *Times*.

Recentemente analisei dezenas de milhares de perfis do Stormfront, em que membros registrados podem inserir sua localização, data de nascimento, interesses e outras informações.

O Stormfront foi fundado em 1995 por Don Black, ex-líder da Ku Klux Klan. Seus "grupos sociais" mais populares são "União de Socialistas Nacionais" e "Fãs e Apoiadores de Adolf Hitler". Ao longo do último ano, de acordo com o Quantcast, aproximadamente 200 mil a 400 mil norte-americanos visitaram o site mensalmente. Um recente relatório emitido pela Southern Poverty Law Center (organização de advocacia sem fins lucrativos dos Estados Unidos especializada em direito civis e litígios de interesse público) relacionou quase mil homicídios nos últimos cinco anos a membros registrados do Stormfront.

Os membros do Stormfront não são quem eu teria adivinhado. Eles tendem a ser jovens, pelo menos de acordo com as datas de nascimento informadas. A idade mais comum que as pessoas ingressam no site é 19 anos. E há quatro vezes mais jovens de 19 anos do que homens de 40 registrados. Usuários de internet e rede social tendem a ser em média jovens, mas não tão jovens.

Os perfis não têm um campo para gênero. Mas analisei os posts e perfis completos de uma amostra aleatória de usuários norte-americanos e ao que parece é possível descobrir o gênero da maioria dos membros: estimo que 30% dos membros do Stormfront são mulheres.

Os estados com maior número de membros *per capita* são Montana, Alaska e Idaho. Estes estados tendem a ser esmagadoramente brancos. Isto significa que crescer com pouca diversidade incentiva o ódio?

Provavelmente não. Mas como esses estados têm uma maior proporção de brancos não judeus, tendem a ter mais membros em potencial de grupos que atacam judeus e não brancos. O porcentual do público-alvo do Stormfront que realmente se filia ao site é de fato mais alto em áreas com mais minorias. Isto é especialmente verdade quando se observa os membros do Stormfront com 18 anos ou menos e que, portanto, não escolhem onde morar.

Entre este grupo etário, a Califórnia, um estado com uma das maiores populações minoritárias, tem uma taxa de filiação 25% maior do que a média nacional.

Um dos grupos sociais mais populares no site é "Em Apoio ao Antissemitismo". O porcentual dos membros que ingressaram neste grupo está positivamente correlacionado à população de judeus do estado. Nova York, o estado com a população judia mais elevada, tem taxa de afiliação *per capita* acima da média neste grupo.

Em 2001, Dna88 ingressou no Stormfront, descrevendo a si mesmo como "um programador web de 30 anos de boa aparência, racialmente consciente que vive na cidade de Judeu York". Nos quatro meses seguintes ele escreveu mais de duzentos posts, como "Crimes de Judeus contra a Humanidade" e "Dinheiro Sujo dos Judeus", direcionando pessoas para um site, jewwatch.com, que alega ser uma "biblioteca especializada" em "criminalidade sionista".

Os membros do Stormfront reclamam sobre as minorias que falam idiomas diferentes e cometem crimes. Mas o que descobri de mais interessante foi sobre a concorrência no mercado romântico.

Um homem que se autodenomina William Lyon Mackenzie King, em homenagem a um ex-primeiro ministro do Canadá que certa vez sugeriu que o "Canadá deveria permanecer um país de brancos", escreveu em 2003 que lutou para "conter" seu "ódio" depois de ver uma mulher branca "carregando seu mulatinho horroroso por aí". Em seu perfil, Whitepride26, uma estudante de 41 anos de Los Angeles, diz: "Não gosto de negras, latinas e de algumas asiáticas, especialmente quando os homens as acham mais atraentes do que uma mulher branca."

Certos eventos políticos desempenham um papel. O dia de maior aumento na filiação da história do Stormfront foi, de longe, 5 de novembro de 2008, o dia após a eleição de Barack Obama como presidente. Não houve, porém, um aumento de interesse no Stormfront durante a candidatura de Donald Trump, e houve apenas um pequeno aumento imediatamente depois de sua vitória. Trump desfrutou da onda de nacionalismo branco. Não existem evidências aqui de que ele tenha criado essa onda.

A eleição de Obama levou a uma explosão do movimento de nacionalismo branco. A eleição de Trump parece ser uma resposta deste movimento.

Uma coisa que não parece fazer diferença: a economia. Não houve relação entre os registros mensais de filiação e a taxa de desemprego de um estado. Os estados desproporcionalmente afetados pela Grande Recessão não tiveram um aumento comparativo nas buscas no Google por Stormfront.

Mas talvez o que tenha sido mais interessante — e surpreendente — foram alguns dos tópicos de conversas entre os membros do Stormfront. Eles são semelhantes aos que meus amigos e eu conversamos. Talvez tenha sido ingenuidade minha, mas eu imaginaria nacionalistas habitando um universo diferente daquele de meus amigos e eu. Em vez disso, eles travam longas discussões elogiando *Game of Thrones* e discutindo as vantagens comparativas de sites de encontros na internet, como PlentyOfFish e OkCupid.

E o fator-chave que mostra que os usuários do Stormfront habitam universos parecidos ao de pessoas como eu e meus amigos: a popularidade do *New York Times* entre os usuários do Stormfront. Não é apenas a VikingMaiden88 que gosta de visitar o site do *Times*. O site é popular entre muitos de seus membros. Na verdade, quando comparamos os usuários do Stormfront a pessoas que visitam o site Yahoo News, os membros do Stormfront têm duas vezes mais probabilidade de visitar o nytimes.com.

Os membros dos sites de ódio lendo o tão liberal nytimes.com? Como isto seria possível? Se um número substancial de membros de Stormfront obtém suas notícias no nytimes.com, isto significa que a sabedoria convencional sobre o nacionalismo branco está errada. Também significa que a sabedoria convencional sobre como a internet funciona está errada.

A VERDADE SOBRE A INTERNET

A internet, quase todo mundo concorda, divide os norte-americanos, fazendo com que muitas pessoas se escondam em sites com pessoas que gostem delas. Cass Sunstein, da Harvard Law School, descreveu a situação como: “Nosso mercado de comunicações está rapidamente se movendo [na direção de uma situação em que] as pessoas se restringem àqueles que têm os mesmos pontos de vista que os seus — liberais assistem a e leem principalmente ou apenas liberais; moderados, apenas moderados; conservadores, somente conservadores, neonazis, só neonazis.”

Esta visão faz sentido. Afinal, a internet nos oferece um número virtualmente iluminado de opções a partir das quais consumir notícias. Posso ler o que eu quiser. Você pode ler o que quiser. VikingMaiden88 pode ler o que ela quiser. E as pessoas, se puderem decidir por si mesmas, tendem a buscar pontos de vistas que confirmem suas crenças. Assim, certamente, a internet cria extrema segregação política.

Existe um problema com essa visão padrão. Os dados nos dizem que isto simplesmente não é verdade.

As provas contra esta sabedoria convencional vêm de um estudo de 2011 de Matt Gentzkow e Jesse Shapiro, dois economistas cujo trabalho discutimos anteriormente.

Gentzkow e Shapiro coletaram dados do comportamento de navegação de uma grande amostra de norte-americanos. O conjunto de dados deles também incluiu a ideologia — autorrelatada — de seus objetos de estudo: se a pessoa se considerava mais liberal ou conservadora. Eles usaram esses dados para medir a segregação política na internet.

Como? Realizaram um interessante experimento de pensamento.

Suponha que você coletara uma amostragem aleatória de dois norte-americanos, ambos visitando o mesmo site de notícias. Qual é a probabilidade de um deles ser liberal e o outro, conservador? Com que frequência, em outras palavras, liberais e conservadores se “encontram” em sites de notícias?

Para refletir um pouco mais sobre a questão, suponha que liberais e conservadores na internet nunca obtenham notícias online do mesmo lugar. Em outras palavras, que liberais visitem exclusivamente sites liberais, conservadores, exclusivamente sites conservadores. Se este fosse o caso, a probabilidade de dois norte-americanos visitando um determinado site de notícias com visões políticas opostas seria 0%. A internet seria perfeitamente *segregada*. Liberais e conservadores nunca se misturariam.

Suponha, em contrapartida, que liberais e conservadores não difiram no modo como obtêm suas notícias. Em outras palavras, um liberal e um conservador eram igualmente propensos a visitar qualquer site específico de notícias. Se este fosse o caso, a probabilidade de dois norte-americanos em um determinado site de notícias com

visões políticas opostas seria de aproximadamente 50%. A internet estaria perfeitamente *agregada*. Liberais e conservadores se misturariam com perfeição.

Então, o que os dados nos mostram? Nos Estados Unidos, de acordo com Gentzkow e Shapiro, a probabilidade de que duas pessoas que visitam o mesmo site de notícias tenham visões políticas diferentes é de 45%. Em outras palavras, a internet está muito mais perto da perfeita agregação do que da total segregação. Liberais e conservadores se “encontram” na web o tempo todo.

O que realmente coloca a falta de segregação na internet em perspectiva é compará-la com a segregação em outras partes de nossas vidas. Gentzkow e Shapiro poderiam repetir a análise para diversas interações offline. Quais as probabilidades de dois membros da mesma família terem visões políticas diferentes? Dois vizinhos? Dois colegas? Dois amigos?

Usando os dados da General Social Survey, Gentzkow e Shapiro descobriram que todos esses números eram menores do que as chances de duas pessoas no mesmo site de notícias terem visões políticas diferentes.

PROBABILIDADE DE QUE ALGUÉM QUE VOCÊ CONHEÇA TENHA VISÃO POLÍTICA OPOSTA

Em um site de notícias	45,2%
Colega de trabalho	41,6%
Vizinho offline	40,3%
Familiar	37,0%
Amigo	34,7%

Em outras palavras, você tem mais probabilidade de encontrar alguém com visões opostas online do que offline.

Por que a internet não é mais segregada? Há dois fatores que limitam a segregação política na internet.

Primeiro, o que é um tanto surpreendente, a indústria de notícias da internet é dominada por alguns poucos sites gigantescos. Normalmente pensamos na internet como algo que atrai movimentos marginais. Na verdade, há sites para todo mundo, não importa o ponto de vista. Há lugar para defensores pró-armas e antiarmas, direitos dos fumantes de charutos e defensores de uma moeda forte, anarquistas e nacionalistas brancos. Mas esses sites juntos respondem por uma pequena fração do tráfego de notícias da internet. Na verdade, em 2009, quatro sites — Yahoo News, AOL News, msnbc.com e cnn.com — obtiveram mais da metade das visualizações de notícias. O Yahoo News permanece o site mais popular entre os norte-americanos, com perto de 90 milhões de visitantes únicos mensalmente — ou 600 vezes o público do Stormfront. Sites de mídia de massa como o Yahoo News atraem um público abrangente e politicamente diverso.

A segunda razão pela qual a internet não é tão segregada é que muitas pessoas com fortes opiniões políticas visitam sites de pontos de vista opostos, nem que seja apenas para ficar irritado e discutir. Viciados em política não se limitam apenas a sites que defendam suas posições. Alguém que visita o thinkprogress.org e o moveon.org — dois sites extremamente liberais — tem mais chance do que o usuário médio da internet de visitar o foxnews.com, de inclinação direitista. Alguém que visite o rushlimbaugh.com ou o glennbeck.com — extremamente conservadores — tem mais probabilidade do que o usuário médio de visitar o nytimes.com, mais liberal.

O estudo de Gentzkow e Shapiro foi baseado nos dados entre 2004–2009, relativamente cedo na história da internet. Poderia a internet ter se tornado mais compartimentalizada desde então? A mídia social e, em especial, o Facebook mudaram a conclusão deles? Claramente, se nossos amigos tendem a compartilhar de nossas visões políticas, o aumento da mídia social deveria significar um aumento nas câmaras de eco. Certo?

Novamente, a história não é simples. Enquanto é verdade que os amigos de uma pessoa no Facebook são mais propensos a compartilhar de suas visões políticas, uma equipe de cientistas de dados — Eytan Bakshy, Solomon Messing e Lada Adamic — descobriu que uma quantidade surpreendente de informações que as pessoas obtêm no Facebook vem de pessoas com pontos de vista opostos.

Como pode? Nossos amigos não tendem a partilhar de nossas visões políticas? De fato, sim. Mas há um motivo crucial para que o Facebook possa levar a uma discussão política mais diversificada do que a socialização offline. Pessoas, em média, têm substancialmente mais amigos no Facebook do que offline. E estes laços de amizades frágeis facilitados pelo Facebook têm maior probabilidade de envolver pessoas com visões políticas opostas.

Em outras palavras, o Facebook nos expõe a conexões sociais frágeis — o conhecido dos tempos de colégio, o primo maluco de terceiro grau, o amigo do amigo do amigo que você meio que talvez conheça. Estas são pessoas com quem você provavelmente nunca sairá para jogar boliche ou convidará para um churrasco. Não chamará para

festas. Mas você os adiciona como amigos no Facebook. E visualiza os links para os artigos postados por eles com visões que nunca consideraria.

Resumindo, a internet de fato aproxima pessoas de diferentes visões políticas. A liberal média passa a manhã com seu marido liberal e seus filhos liberais; as tardes, com seus colegas de trabalho liberais; no caminho de casa só vê carros com adesivos liberais; as noites, com suas colegas de yoga liberais. Quando chega em casa lê alguns comentários conservadores na cnn.com ou visualiza um link no Facebook de uma Republicana, sua colega do ensino médio; esta pode ser sua maior exposição conservadora do dia.

Provavelmente nunca encontrei nacionalistas brancos em minha cafeteria preferida no Brooklyn. Mas VikingMaiden88 e eu frequentamos o site do *New York Times*.

A VERDADE SOBRE O ABUSO INFANTIL E O ABORTO

A internet pode nos dar novas perspectivas não só sobre atitudes, mas também sobre comportamentos perturbadores. De fato, os dados do Google são eficientes em nos alertar sobre crises ignoradas pelas fontes convencionais. As pessoas, afinal, recorrem ao Google quando estão com problemas.

Pense no abuso infantil durante a Grande Recessão.

Quando este gigantesco revés econômico começou, no fim de 2007, muitos especialistas estavam naturalmente preocupados com o efeito que teria sobre as crianças. Afinal, muitos pais estariam estressados e deprimidos, e estes são os principais fatores de risco para maus-tratos. O abuso infantil poderia disparar.

Então os dados oficiais foram divulgados, e parecia que a preocupação era infundada. Os órgãos de assistência social infantil reportaram que estavam recebendo menos casos de abuso. Além disto, essas quedas eram menores nos estados mais atingidos pela recessão. “As previsões sombrias não se tornaram realidade”, disse Richard Gelles, um especialista em bem-estar infantil da Universidade da Pensilvânia, para a Associated Press, em 2011. Sim, por mais contraintuitivo que possa ser, o abuso infantil pareceu despencar durante a recessão.

Mas o abuso infantil realmente diminuiu com tantos adultos sem emprego e extremamente estressados? Não consegui acreditar. Então, recorri aos dados do Google.

Ao que parece, algumas crianças fizeram buscas trágicas e comoventes no Google — tais como “minha mãe me bateu” ou “meu pai me bateu”. E estas buscas apresentam um quadro diferente — e cruel — do que aconteceu durante a época. O número de buscas assim disparou durante a Grande Recessão, acompanhando de perto a taxa de desemprego.

Veja o que acho que aconteceu: o número de denúncias de casos de abuso infantil diminuiu, não os abusos em si. Afinal, estima-se que apenas um pequeno porcentual de casos de abuso infantil é de fato reportado para as autoridades. E durante a recessão, muitas pessoas que tendem a denunciar (professores e policiais, por exemplo) e lidar com casos de abuso infantil (assistentes sociais) têm maior probabilidade de estar sobrecarregadas ou desempregadas.

Havia muitas histórias durante o revés econômico de pessoas tentando denunciar potenciais casos, que precisam enfrentar longas esperas e acabam desistindo.

Na verdade, há mais provas, desta vez não do Google, de que o abuso infantil de fato subiu durante a crise. Quando uma criança morre em razão de abuso ou negligência, isso precisa ser reportado. Essas mortes, embora raras, aumentaram em estados atingidos de modo mais severo pela recessão.

E há algumas evidências no Google do aumento das suspeitas de abuso infantil em áreas muito atingidas. Ao se fazer o controle nas taxas pré-recessão e nas tendências nacionais, os estados que comparativamente sofreram mais tiveram taxas de buscas aumentadas para abuso e negligência infantis. Para cada ponto porcentual de aumento na taxa de desemprego, houve um aumento de 3% associado na taxa de busca para “abuso infantil” ou “negligência infantil”. Presumivelmente, a maioria dessas pessoas nunca relatou abuso, já que esses estados tiveram as maiores quedas nos relatos.

As buscas feitas por crianças em sofrimento aumentaram. A taxa de morte de crianças sofreu um pico. As buscas realizadas por pessoas suspeitando de abuso subiram em estados muito atingidos. Mas os relatos de casos diminuíram. Uma crise parece fazer com que mais crianças contem ao Google que seus pais estão lhe batendo ou espancando, e mais pessoas suspeitam ter presenciado abuso. Mas os órgãos sobrecarregados são capazes de lidar com menos casos.

Acredito que é seguro dizer que a Grande Recessão piorou o abuso infantil, embora as medições tradicionais não mostrem isto.

Toda vez que desconfio que as pessoas estejam sofrendo em segredo agora, recorro aos dados do Google. Um dos potenciais benefícios destes novos dados, e de saber como os interpretar, é a possibilidade de ajudar pessoas

em situação de vulnerabilidade, que de outra forma seriam ignoradas pelas autoridades.

Então, quando a Suprema Corte dos Estados Unidos examinou, recentemente, os efeitos das leis que dificultam o acesso ao aborto, recorri aos dados de busca. Suspeitei que mulheres afetadas pela legislação procurariam por maneiras não oficiais de interromper uma gravidez. E de fato, elas procuram. E essas buscas foram mais frequentes em estados que aprovaram leis restringindo o aborto.

Os dados de busca aqui são tanto úteis quanto preocupantes.

Em 2015, nos Estados Unidos, houve mais de 700 mil buscas procurando por abortos autoinduzidos. Por comparação, houve cerca de 3,4 milhões de buscas por clínicas de aborto naquele ano. Isto sugere que um porcentual significativo de mulheres considerando um aborto contemplou a ideia de fazê-lo sozinhas.

As mulheres pesquisaram, cerca de 160 mil vezes, formas de obter pílulas abortivas através de canais não oficiais — “comprar pílulas abortivas online” e “pílulas abortivas grátis”. Elas perguntaram ao Google sobre aborto com ervas como salsa ou com vitamina C. Houve cerca de 4 mil buscas procurando por instruções sobre como abortar usando um cabide, incluindo aproximadamente 1.300 pela frase exata “como fazer um aborto com cabide”. Houve também algumas centenas procurando por aborto aplicando água sanitária no útero e golpeando o abdômen.

O que impulsiona o interesse em um aborto autoinduzido? A geografia e o momento das buscas no Google apontam para um provável culpado: quando é difícil conseguir um aborto oficial, as mulheres buscam por abordagens não oficiais.

As taxas de busca por aborto autoinduzido eram bastante estáveis entre 2004 e 2007. Elas começaram a aumentar no final de 2008, coincidindo com a crise financeira e a recessão que a seguiu. As buscas deram um grande salto em 2011, aumentando em 40%. O Instituto Guttmacher, uma organização de direitos reprodutivos, destaca 2011 como o início da recente ação restritiva do país sobre o aborto; 92 disposições legais restringindo o acesso ao aborto foram aprovadas. Ao se comparar com o Canadá, que não sofreu qualquer restrição legal aos direitos reprodutivos, não houve aumentos comparáveis em buscas por abortos autoinduzidos durante o período.

O estado com a taxa de busca no Google mais alta por abortos autoinduzidos é Mississippi, um estado com aproximadamente três milhões de pessoas e, agora, apenas uma clínica de aborto. Oito dos dez estados com taxas de busca mais altas por aborto autoinduzido são considerados pelo Instituto Guttmacher como sendo hostis ou muito hostis ao aborto. Nenhum dos dez estados com taxas de busca mais baixas por aborto estão nestas categorias.

Obviamente, não podemos saber, a partir dos dados do Google, quantas mulheres realmente se submetem a abortos, mas evidências sugerem que um número significativo acaba realmente os fazendo. Uma forma de esclarecer isto é comparar dados de aborto e de nascimentos.

Em 2011, o último ano com dados completos sobre o aborto, as mulheres morando em estados com menos clínicas de aborto fizeram muito menos abortos legais.

Compare os dez estados com mais quantidade de clínicas de aborto *per capita* (uma lista que inclui Nova York e Califórnia) com os dez com menos (uma lista que inclui Mississippi e Oklahoma). Mulheres que moram em estados com menos clínicas tiveram 54% menos abortos legais — uma diferença de onze abortos para cada mil mulheres entre 15 e 44 anos. As mulheres que moram em estados com menos clínicas também tiveram mais nascimentos. Entretanto, a diferença não foi suficiente para compensar o menor número de abortos. Houve seis nascimentos a mais para cada mil mulheres em idade fértil.

Em outras palavras, ao que parece há algumas gravidezes ausentes nos dados em partes do país em que é mais difícil conseguir fazer um aborto. As fontes oficiais não nos dizem o que aconteceu com aqueles cinco nascimentos ausentes para cada mil mulheres em estados em que o acesso ao aborto é mais difícil.

O Google fornece algumas pistas muito boas.

Não podemos confiar cegamente nos dados governamentais. O governo pode nos dizer que o abuso infantil ou o aborto diminuíram, e os políticos, celebrar este feito. Mas os resultados que pensamos que vemos são um produto das falhas nos métodos de coleção de dados. A verdade é bem diferente — e, às vezes, muito mais sombria.

A VERDADE SOBRE SEUS AMIGOS NO FACEBOOK

Este livro é sobre Big Data, em geral. Mas este capítulo enfatiza principalmente as buscas no Google, que, defendo, revelam um mundo secreto muito diferente daquele que pensamos ver. Será que outras fontes de Big Data também são o soro da verdade digital? O fato é que muitas fontes de Big Data, como o Facebook, são com frequência o oposto do soro da verdade digital.

Na mídia social, assim como em pesquisas, não há incentivo para se dizer a verdade. Nas redes sociais, muito mais do que nas pesquisas, há um grande incentivo para transmitir uma boa imagem. Sua presença online não é

anônima, afinal. Você tenta agradar um público e dizer a seus amigos, familiares, colegas, conhecidos e estranhos quem é.

Para ver o quanto os dados da mídia social são tendenciosos, pense na relativa popularidade da *Atlantic*, uma revista mensal respeitada e refinada, versus a *National Enquirer*, uma revista sensacionalista de fofocas. Ambas as publicações têm circulação média similar, vendendo algumas centenas de milhares de cópias. (A *National Enquirer* é semanal, então, na verdade, vende muito mais cópias totais.) Há ainda um número equivalente de buscas no Google para cada revista.

Entretanto, no Facebook, aproximadamente 1,5 milhões de usuários curtem a *Atlantic* ou discutem artigos da *Atlantic* em seus perfis. Apenas cerca de 50 mil curtem a *Enquirer* ou discutem seu conteúdo.

POPULARIDADE DA *ATLANTIC* VERSUS *NATIONAL ENQUIRER* COMPARADA EM DIFERENTES FONTES

Circulação	Aproximadamente 1 *Atlantic* para 1 *National Enquirer*
Buscas no Google	1 *Atlantic* para 1 *National Enquirer*
Curtidas no Facebook	27 *Atlantic* para 1 da *National Enquirer*

Para avaliar a popularidade de uma revista, os dados sobre sua circulação são a verdade nua e crua. Os dados do Google são muito parecidos. E os dados do Facebook são extremamente preconceituosos contra o tabloide fútil, o que os tornam os piores dados para determinar do que as pessoas realmente gostam.

E é assim em relação a preferências de leitura e também na vida. No Facebook, mostramos nosso eu lapidado, não nosso verdadeiro eu. Uso os dados do Facebook neste livro, na verdade neste capítulo, mas sempre com esta limitação em mente.

Para obter uma melhor compreensão do que a rede social omite, vamos voltar por um momento para a pornografia. Primeiro, precisamos lidar com a crença comum de que a internet é dominada por obscenidade. Isto não é verdade. A maior parte do conteúdo da internet é não pornográfico. Por exemplo, dos dez sites mais visitados, nenhum deles é pornô. Assim, a popularidade da pornografia, apesar de enorme, não deve ser superestimada.

Ainda assim, porém, ao analisarmos mais de perto como curtimos e compartilhamos a pornografia, fica claro que Facebook, Instagram e Twitter oferecem apenas uma visão limitada do que é realmente popular na internet. Há grandes subconjuntos da web que operam com gigantesca popularidade, mas pouca presença social.

O vídeo mais popular de todos os tempos, até o momento em que escrevo, é “Gangnam Style”, de Psy, um vídeo bobo de música pop que satiriza a moda coreana. Ele foi visualizado cerca de 2,3 bilhões de vezes apenas no YouTube desde sua estreia, em 2012. E sua popularidade é evidente, não importa em que site você esteja. Ele foi compartilhado em diferentes plataformas de mídia social milhões de vezes.

O vídeo pornô mais popular de todos os tempos pode ser “Great Body, Great Sex, Great Blowjob” [“Corpo Fantástico, Sexo Fantástico, Boquete Fantástico”, em tradução livre]. Ele foi visualizado mais de 80 milhões de vezes. Em outras palavras, para cada trinta visualizações de “Gangnam Style” houve aproximadamente uma de “Great Body, Great Sex, Great Blowjob”. Se a mídia social nos oferecesse uma visão precisa dos vídeos a que as pessoas assistem, “Great Body, Great Sex, Great Blowjob” deveria ter sido postado milhões de vezes. Mas foi compartilhado em mídias sociais apenas algumas dezenas de vezes, e sempre por estrelas pornôs, não usuários comuns. As pessoas claramente não sentem a necessidade de propagandear seus interesses neste vídeo para os amigos.

O Facebook é o soro digital do “me gabar para meus amigos de como minha vida é boa”. No mundo do Facebook, o adulto médio parece ser casado e feliz, tirar férias no Caribe e ler a *Atlantic*. No mundo real, muitas pessoas estão irritadas nas filas de supermercados espiando a *National Enquirer*, ignorando as ligações dos cônjuges, com quem não transam há anos. No mundo do Facebook, as famílias parecem perfeitas. No mundo real, a vida familiar é complicada. Pode ser tão complexa a ponto de ocasionalmente um pequeno número de pessoas se arrepender de ter filhos. No mundo do Facebook, parece que todo jovem adulto está em uma festa maravilhosa no sábado à noite. No mundo real, a maioria está em casa sozinho, assistindo a maratonas de séries no Netflix. No mundo do Facebook, a namorada posta 26 fotos felizes de sua escapada romântica com o namorado. No mundo real, imediatamente depois de postar as fotos, pesquisa no Google: “Meu namorado não quer transar comigo.” E, talvez ao mesmo tempo, o namorado assista a “Great Body, Great Sex, Great Blowjob”.

VERDADES DIGITAIS	MENTIRAS DIGITAIS
• Buscas	• Posts na mídia social
• Visualizações	• Curtidas na mídia social
• Cliques	• Perfis de namoros
• Rolagens de tela	

A VERDADE SOBRE SEUS CLIENTES

Na manhã do dia 5 de setembro de 2006, o Facebook introduziu uma grande atualização em sua página inicial. As primeiras versões do Facebook somente permitiam aos usuários clicar nos perfis de seus amigos para saber o que faziam. O site, considerado um grande sucesso, tinha, na época, 9,4 milhões de usuários.

Mas depois de meses de trabalho árduo, os engenheiros criaram algo chamado "feed de notícias", que oferece aos usuários atualizações das atividades de todos os seus amigos.

Imediatamente, usuários reportaram odiar o feed de notícias. Ben Parr, graduando da Universidade de Northwestern, criou um grupo, "Alunos contra o feed de notícias do Facebook". Ele disse: "O feed de notícias é simplesmente sinistro, bisbilhotice demais, um recurso que precisa ser removido." Em poucos dias, o grupo tinha 700 mil membros ecoando o sentimento de Parr. Um calouro da Universidade de Michigan disse ao *Michigan Daily*: "Estou realmente apavorado com este novo Facebook. Ele faz com que eu me sinta um perseguidor."

David Kirkpatrick conta este episódio em sua narração autorizada da história do site, *O Efeito Facebook: Os Bastidores da História da Empresa que Conecta o Mundo*. Ele apelidou a introdução do feed de notícias de "a maior crise jamais enfrentada pelo Facebook". Mas Kirkpatrick relata que, quando entrevistou Mark Zuckerberg, cofundador e líder da empresa de crescimento acelerado, o CEO estava inabalável.

O motivo? Zuckerberg tinha acesso ao soro da verdade digital: inúmero cliques e visitas de pessoas ao Facebook. Como escreve Kirkpatrick:

> Zuckerberg de fato sabia que as pessoas gostaram do feed de notícias, não importa o que diziam nos grupos. Ele tinha os dados para provar. As pessoas estavam passando, em média, mais tempo no Facebook do que antes do lançamento do feed. E faziam mais coisas — muito mais. Em agosto, os usuários visualizaram 12 bilhões de páginas através do site. Mas, em outubro, com o feed funcionando, visualizaram 22 bilhões.

E isso não era toda a evidência à disposição de Zuckerberg. Até a popularidade viral do grupo contra o feed de notícias era prova do seu poder. O grupo foi capaz de crescer tão rápido exatamente porque um grande número de pessoas descobriu que seus amigos haviam ingressado no grupo — e isso só foi possível através do feed de notícias.

Em outras palavras, ao mesmo tempo que as pessoas se juntavam ao grupo em um grande alvoroço público do quanto estavam descontentes em ver todos os detalhes das vidas de seus amigos no Facebook, continuavam retornando para saber mais detalhes sobre a vida dos amigos. O feed de notícias ficou. O Facebook agora tinha mais de um bilhão de usuários ativos diariamente.

Em seu livro *De Zero a Um*, Peter Thiel, um dos investidores iniciais no Facebook, disse que ótimos negócios são construídos sobre segredos, sejam da natureza ou das pessoas. Jeff Seder, como mencionei no Capítulo 3, descobriu o segredo natural de que o tamanho do ventrículo esquerdo era capaz de predizer o desempenho do cavalo. O Google descobriu o segredo natural do poder da informação contida nos links.

Thiel define "segredos das pessoas" como "coisas que não sabem sobre si mesmas ou que escondem porque não querem que os outros saibam". Estes tipos de negócio, em outras palavras, são construídos sobre a mentira das pessoas.

Você pode argumentar que tudo no Facebook é fundamentado em um desagradável segredo das pessoas que Zuckerberg descobriu em Harvard. Zuckerberg, no início de seu segundo ano, criou um site para seus colegas estudantes chamado Facemash. Baseado em um site chamado "Am I Hot or Not?" ["Sou Sexy ou Não?", em tradução livre], o Facemash apresentava fotos de dois alunos de Harvard e depois pedia que os outros julgassem quem era mais bonito.

O site foi recebido com ultraje. O *Harvard Crimson*, em um editorial, acusou o jovem Zuckerberg de "alimentar o pior lado" das pessoas. Grupos hispânicos e afro-americanos o acusaram de sexismo e racismo. Ainda assim, antes de os administradores de Harvard desligarem o acesso de Zuckerberg à internet — apenas algumas

poucas horas depois que o site fora fundado — 450 pessoas tinham visualizado e votado 22 mil vezes em diferentes imagens. Zuckerberg descobriu um importante segredo: as pessoas podem até alegar estarem furiosas, podem condenar algo como sendo de mau gosto, e, ainda assim, clicam.

E descobriu mais uma coisa: para todas as alegações de seriedade, responsabilidade e respeito pela privacidade alheia, as pessoas, mesmo os alunos de Harvard, tinham um grande interesse em avaliar a aparência das pessoas. As visualizações e os votos demonstraram isso. E mais tarde — já que o Facemash se provou muito controverso — ele pegou este conhecimento sobre o interesse das pessoas em fatos superficiais da vida de pessoas que mal conhecem e o empregou na empresa mais bem-sucedida de sua geração.

A Netflix aprendeu uma lição semelhante logo no início de sua vida: não confie no que as pessoas lhe dizem; confie no que fazem.

Originalmente, a Netflix permitia aos usuários criar uma lista de filmes a que queriam assistir no futuro, mas que não tinham tempo no momento. Dessa forma, quando tivessem mais tempo, a Netflix lhes recordaria sobre aqueles filmes.

Entretanto, a Netflix percebeu algo estranho nos dados. Os usuários enchiam suas listas com inúmeros filmes. Mas dias depois, quando eram notificados, raramente clicavam no link.

Qual era o problema? Pergunte aos usuários a quais filmes planejam assistir dentro de alguns dias e eles lotarão a lista com filmes cultos e inoperacionais, como documentários em preto e branco sobre a Segunda Guerra Mundial ou filmes “cabeça” estrangeiros. Alguns dias depois, porém, eles preferem assistir aos mesmos filmes a que costumam assistir: comédias bobas ou filmes românticos. As pessoas mentiam para si mesmas reiteradamente.

Diante desta disparidade, a Netflix parou de perguntar aos usuários ao que queriam assistir no futuro e começou a criar um modelo baseado em milhões de cliques e visualizações de clientes semelhantes. A empresa começou a oferecer aos usuários listas de filmes sugeridos com base não no que alegavam gostar, mas no que os dados disseram a que gostariam de assistir. O resultado: os clientes visitavam o site da Netflix com mais frequência e assistiam a mais filmes.

“Os algoritmos conhecem você melhor do que você mesmo”, diz Xavier Amatriain, ex-cientista de dados da Netflix.

O VALOR DE IGNORAR O QUE AS PESSOAS DIZEM A VOCÊ

O QUE AS PESSOAS DIZEM	REALIDADE	CONSEQUÊNCIA...
Não querem "perseguir" os amigos.	Poucas coisas no mundo as pessoas querem mais do que acompanhar e julgar seus amigos.	Mark Zuckerberg, cofundador do Facebook, vale US$55,2 bilhões.
Não querem comprar produtos feitos em fábricas clandestinas que exploram os trabalhadores.	Elas compram produtos bons com "preços razoáveis".	Phil Knight, cofundador da Nike, vale US$25,4 bilhões.
Querem ouvir notícias pela manhã.	Elas querem ouvir sobre anãos transando com estrelas pornôs pela manhã.	Howard Stern vale US$500 milhões.
Não têm interesse em ler sobre sadomasoquismo, *bondage* e Dominação.	Elas querem ler sobre relacionamento BDSM entre uma jovem recém-formada e um magnata dos negócios.	***50 Tons de Cinza*** vendeu 125 milhões de cópias mundialmente.
Querem que os políticos definam suas posições políticas.	Elas querem que os políticos as poupem dos detalhes, mas que pareçam fortes e confiantes.	Donald Trump.

SERÁ QUE CONSEGUIMOS ENCARAR A VERDADE?

Você pode achar algumas partes deste capítulo deprimentes. O soro da verdade digital revelou um interesse permanente em julgar as pessoas com base em sua aparência; a contínua existência de milhões de homens gays no armário; um significativo porcentual de mulheres fantasiando sobre estupro; o ódio disseminado contra afro-americanos; uma crise oculta de abuso infantil e aborto autoinduzido e um surto de violento ódio islamofóbico que apenas piora quando o presidente apela por tolerância. Não são exatamente informações animadoras. Frequentemente, depois de palestras sobre minha pesquisa, as pessoas vêm até mim e dizem: "Seth, tudo isso é muito interessante. Mas também tão deprimente."

Não posso fingir que não exista essa obscuridade em alguns dados. Se as pessoas nos dizem continuamente o que pensam que queremos ouvir, constantemente ouviremos coisas mais reconfortantes que a verdade. O soro da verdade digital, em média, nos mostrará que o mundo é pior do que pensávamos.

Precisamos saber disso? Descobrir sobre as buscas no Google, os dados da pornografia e quem clica no quê pode não fazer com que você pense: "Isto é fantástico. Podemos entender quem realmente somos." Em vez disso, você pode pensar: "Isto é horroroso. Podemos entender quem realmente somos."

Mas a verdade ajuda — e não apenas Mark Zuckerberg ou outros que buscam atrair cliques ou clientes. Há pelo menos três formas de esse conhecimento melhorar nossa vida.

Primeiro, pode ser reconfortante saber que não está sozinho em suas inseguranças e comportamentos embaraçosos. Pode ser agradável saber que outras pessoas também se sentem inseguras quanto ao próprio corpo. Provavelmente é bom para muitas pessoas — especialmente aquelas que não estão fazendo muito sexo — saber que o mundo todo não transa como coelhos. E pode ser valioso para um garoto de ensino médio no Mississippi com uma queda pelo zagueiro do time saber que apesar do baixo número de homens declaradamente gays a seu redor, existem muitos outros que sentem o mesmo tipo de atração.

Há outra área — uma que ainda não discuti — em que as buscas no Google podem ajudar a lhe mostrar que você não está só. Quando jovem, um professor pode ter lhe dito que, se você tem uma dúvida, deve levantar a mão e perguntar, pois se está confuso, outros também estão. Se for ao menos um pouco parecido comigo, você ignorou os conselhos de seu professor e permaneceu em silêncio, com medo de abrir a boca. Suas perguntas eram estúpidas demais, você pensava; as de todo mundo eram sempre mais profundas. Os dados anônimos e reunidos do Google nos dizem de uma vez por todas o quanto nossos professores estavam certos. Muitas questões básicas, menos profundas, também espreitam as mentes das outras pessoas.

Pense nas principais perguntas que norte-americanos fizeram durante o discurso sobre o Estado da União de 2014 de Obama. (Veja foto colorida no final do livro.)

**VOCÊ NÃO É O ÚNICO IMAGINANDO:
AS PERGUNTAS MAIS PESQUISADAS NO GOOGLE DURANTE O DISCURSO SOBRE O ESTADO DA UNIÃO**

Qual é a idade de Obama?
Quem está sentado ao lado de Biden?
Por que Boehner está usando uma gravata verde?
Por que Boehner é cor de laranja?

Agora, você pode ler estas perguntas e pensar o quanto denigrem nossa democracia. Preocupar-se mais com a cor da gravata de alguém ou com seu tom de pele do que com o conteúdo do discurso do presidente não reflete uma boa imagem do país. Não saber quem é John Boehner, então presidente da câmara dos representantes, também não diz grande coisa sobre nosso engajamento político.

Prefiro pensar nestas perguntas como demonstração do conhecimento de nossos professores. Estes são os tipos de perguntas que as pessoas normalmente não fazem, pois parecem tolas demais. Mas muitos querem saber — e as pesquisam no Google.

Na verdade, penso que o Big Data nos dá uma versão do século XXI da famosa citação: “Nunca compare seu interior com o exterior dos outros.”

Uma atualização do Big Data pode ser: “Nunca compare suas buscas no Google com os posts nas redes sociais dos outros.”

Compare, por exemplo, a forma que as pessoas descrevem seus maridos na mídia social e nas buscas anônimas.

PRINCIPAIS MANEIRAS QUE AS PESSOAS DESCREVEM SEUS MARIDOS

POSTS NAS REDES SOCIAIS	BUSCAS
o melhor	gay
meu melhor amigo	um babaca
maravilhoso	maravilhoso
sensacional	irritante
muito fofo	maldoso

Como vemos os posts das outras pessoas nas redes sociais, mas não suas buscas, tendemos a exagerar no número de mulheres que pensam que seus maridos são “o melhor”, “sensacional” e “muito fofo”.* Tendemos a minimizar quantas pessoas pensam que seus maridos são “um babaca”, “irritante” e “maldoso”. Ao analisar dados anônimos e agrupados, entendemos que não somos os únicos que acham o casamento e a vida difíceis. Podemos aprender a parar de comparar nossas buscas aos posts das outras pessoas na mídia social.

O segundo benefício do soro da verdade digital é que ele nos alerta sobre pessoas que estão sofrendo. A Campanha dos Direitos Humanos me pediu que trabalhasse com eles para ajudar a instruir os homens de certos estados sobre a possibilidade de sair do armário. Eles estão tentando usar os dados de busca agrupados e anônimos do Google para os ajudar a decidir como empregar melhor suas energias. Do mesmo modo, o serviço de assistência social infantil me procurou para descobrir em que partes do país pode haver muito mais abuso infantil do que indicam os registros.

Um tópico surpreendente sobre o qual fui procurado: odores vaginais. Quando escrevi sobre isto pela primeira vez no *New York Times*, fui irônico. A matéria provocou risos em mim e em outras pessoas.

Entretanto, quando mais tarde explorei os painéis de mensagens que surgem quando alguém faz essas buscas, percebi que incluíam inúmeros posts de garotas jovens convencidas de que suas vidas estavam arruinadas em razão da ansiedade sobre os odores vaginais. Não é piada. Especialistas em educação sexual me procuraram perguntando como poderiam incorporar alguns de meus dados da melhor forma para reduzir a paranoia entre aquelas jovens.

Embora me sinta um tanto fora de minha área nestas questões, eles estavam bastante sérios, e acredito que a ciência de dados possa ajudar.

O derradeiro — e, acredito, o mais poderoso — valor deste soro da verdade digital é de fato sua habilidade de nos levar de problemas a soluções. Com mais compreensão, podemos descobrir maneiras de reduzir o suprimento mundial de atitudes desagradáveis.

Vamos voltar ao discurso de Obama sobre islamofobia. Recorde que toda vez que Obama argumentava que as pessoas deveriam respeitar mais os muçulmanos, as mesmas pessoas que tentava alcançar se tornavam mais enfurecidas.

As buscas no Google, porém, revelaram que houve um trecho do discurso que disparou o tipo de resposta que o então presidente queria. Ele disse: “Muçulmanos norte-americanos são nossos amigos e nossos vizinhos, nossos colegas de trabalho, nossos heróis do esporte e, sim, são nossos homens e mulheres fardados, dispostos a morrer na defesa de nosso país.”

Depois destas palavras, pela primeira vez em mais de um ano, os substantivos mais pesquisados depois de “muçulmano” não foram “terroristas”, “extremistas” ou “refugiados”. Foi “atletas”, seguido de “soldados”. E, na verdade, “atletas” manteve a primeira posição por um dia inteiro depois.

Quando falamos para pessoas com raiva, os dados de busca sugerem que sua fúria pode aumentar. Mas provocar sutilmente a curiosidade das pessoas, oferecendo novas informações e imagens do grupo que incitam o ódio direciona seus pensamentos para caminhos mais positivos.

Dois meses depois do discurso original, Obama proferiu outro discurso televisionado sobre a islamofobia, desta vez em uma mesquita. Talvez alguém no gabinete presidencial tenha lido minha coluna e de Solta no *Times*, que discutiu o que havia e o que não havia funcionado, pois o conteúdo deste discurso foi notavelmente diferente.

Obama passou pouco tempo insistindo no valor da tolerância. Em vez disso, focou provocar a curiosidade das pessoas e mudar suas percepções sobre os muçulmanos norte-americanos. Muitos dos escravos vindos da África eram muçulmanos, nos disse Obama; Thomas Jefferson e John Adams tinham suas próprias cópias do Corão; a primeira mesquita nos Estados Unidos foi na Dakota do Norte; um muçulmano norte-americano projetou arranha-céus em Chicago. Obama novamente falou de muçulmanos atletas e membros das forças armadas, mas também falou de policiais, bombeiros, professores e médicos muçulmanos.

E minhas análises das buscas no Google sugerem que esse discurso foi mais bem-sucedido do que o anterior. Muitas das buscas cheias de ódio e raiva contra os muçulmanos despencaram nas horas seguintes ao discurso do ex-presidente.

Existem outras maneiras potenciais de usar os dados de busca para aprender o que causa, ou reduz, o ódio. Por exemplo, podemos analisar como as buscas racistas mudam após um zagueiro negro ser contratado por um time ou como as buscas sexistas se transformam depois que uma mulher é eleita para um cargo público. Podemos ver como o racismo responde ao policiamento comunitário ou como o sexismo reage após novas leis de abuso sexual.

Aprender sobre nossos preconceitos subconscientes também pode ser útil. Por exemplo, podemos tentar fazer um esforço extra para estimular as mentes de meninas e mostrar menos preocupação com sua aparência. Os dados de busca no Google e outras fontes de verdades na internet nos oferecem um olhar inédito nos cantos mais sombrios da psique humana. Isso às vezes é, eu admito, difícil de encarar. Mas também pode ser empoderador. Podemos usar os dados para combater a escuridão. Coletar dados substanciais sobre os problemas do mundo é o primeiro passo para os corrigir.

* Outro motivo para mentir é simplesmente atrapalhar a pesquisa. Esse é um grande problema para qualquer pesquisa envolvendo adolescentes, essencialmente dificultando nossa capacidade de entender este grupo etário. Pesquisadores originalmente descobriram uma correlação entre um

adolescente adotado e uma variedade de comportamentos negativos, tais como uso de drogas, consumo de álcool e matar aula. Em uma pesquisa subsequente, eles descobriram que essa correlação podia ser inteiramente explicada pelas informações dadas por 19% dos adolescentes que, na verdade, não eram adotados. Uma pesquisa subsequente revelou que uma porcentagem significativa dos adolescentes informa nas pesquisas ter mais de dois metros de altura, pesar mais de duzentos quilos e ter três filhos. Uma pesquisa mostrou que 99% dos alunos que relataram possuir uma prótese de braço ou perna estavam brincando.

* Algumas pessoas podem achar ofensiva minha associação da predileção de homens por Judy Garland com a preferência por fazer sexo com homens, mesmo de brincadeira. E certamente não tenho a menor intenção de sugerir isso — ou mesmo que a maioria dos homens gays tenha fascínio por divas. Mas os dados de busca demonstram que há um pouco de verdade neste estereótipo. Estimo que um homem que busca por informação sobre Judy Garland tem três vezes mais probabilidade de buscar por pornografia gay do que hétero. Alguns estereótipos, o Big Data nos mostra, são verdadeiros.

* Acho que estes dados também trazem implicações para uma estratégia de encontros ideais. Claramente, alguém poderia se expor, receber muitas rejeições e não as levar para o lado pessoal. Este processo permitiria eventualmente encontrar um parceiro mais atraído por alguém como você. Não importa sua aparência, essas pessoas são reais. Confie em mim.

* Eu queria chamar este livro de *Meu Pênis É Grande? O que as buscas no Google nos ensinam sobre a natureza humana*, mas meu editor me avisou que seria difícil de vender, que as pessoas poderiam ficar constrangidas demais em comprar um livro com este título em uma livraria de aeroporto. Você concorda?

* Para testar mais a hipótese que os pais tratam os filhos de gêneros diferentes de modo distinto, estou trabalhando na obtenção de dados de sites sobre cuidados maternos e paternos. Eles incluem um número muito maior de pais do que aqueles que fazem estas buscas particulares e específicas.

* Analisei os dados do Twitter. Agradeço a Emma Pierson pela ajuda na obtenção destes dados. Não inclui descritores do que os maridos fazem, o que é predominante nas redes sociais, mas que não tem sentido na busca. Mesmo estes descritores tendem em direções favoráveis. As principais formas de descrever o que um marido está fazendo no momento na rede social são “trabalhando” e “cozinhando”.

5

AJUSTANDO O FOCO

Meu irmão, Noah, é quatro anos mais jovem que eu. Muitas pessoas, quando nos conhecem, nos acham incrivelmente parecidos. Ambos falamos alto, estamos ficando careca da mesma maneira e temos enorme dificuldade em manter nossos apartamentos arrumados.

Mas há diferenças: eu sou econômico. Noah só compra o melhor. Eu adoro Leonard Cohen e Bob Dylan. Noah, Cake e Beck.

Talvez a diferença mais notável entre nós seja nossa atitude em relação ao beisebol. Eu sou obcecado por beisebol e, em especial, meu amor pelo New York Mets sempre foi a principal parte de minha identidade. Noah acha beisebol insuportavelmente chato, e sua aversão pelo esporte sempre foi a principal parte de sua identidade.*

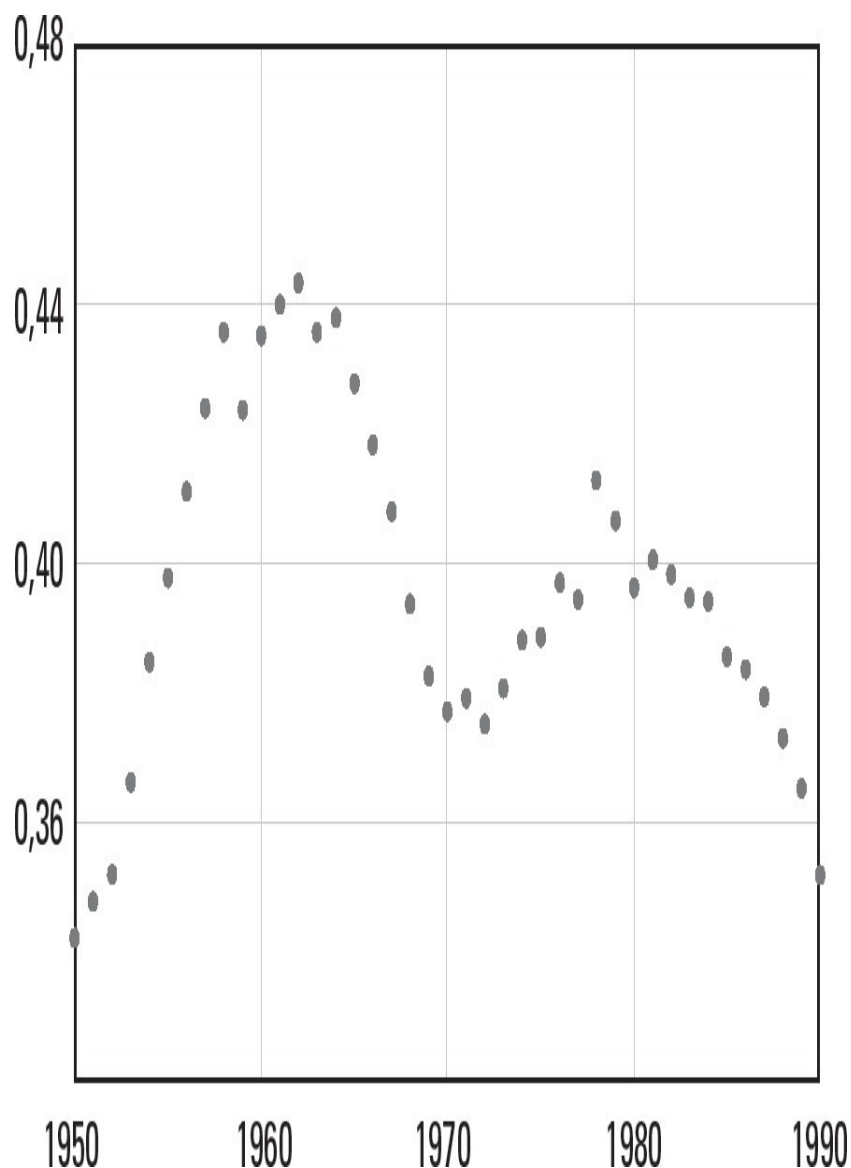
Seth Stephens-Davidowitz Beisebófilo

Noah Stephens-Davidowitz Beisebofóbico

Como podem dois caras com genes parecidos, criados pelos mesmos pais, na mesma cidade, ter sentimentos tão opostos em relação ao beisebol? O que determina os adultos que nos tornaremos? E principalmente, o que há de *errado* com Noah? Existe um campo em crescimento dentro da psicologia desenvolvimental que minera gigantescas bases de dados de adultos e os correlaciona com eventos-chave na infância. Ele nos ajuda a lidar com essa e outras questões análogas. Podemos usar esse crescente uso do Big Data para responder questões psicológicas de Big Psych.

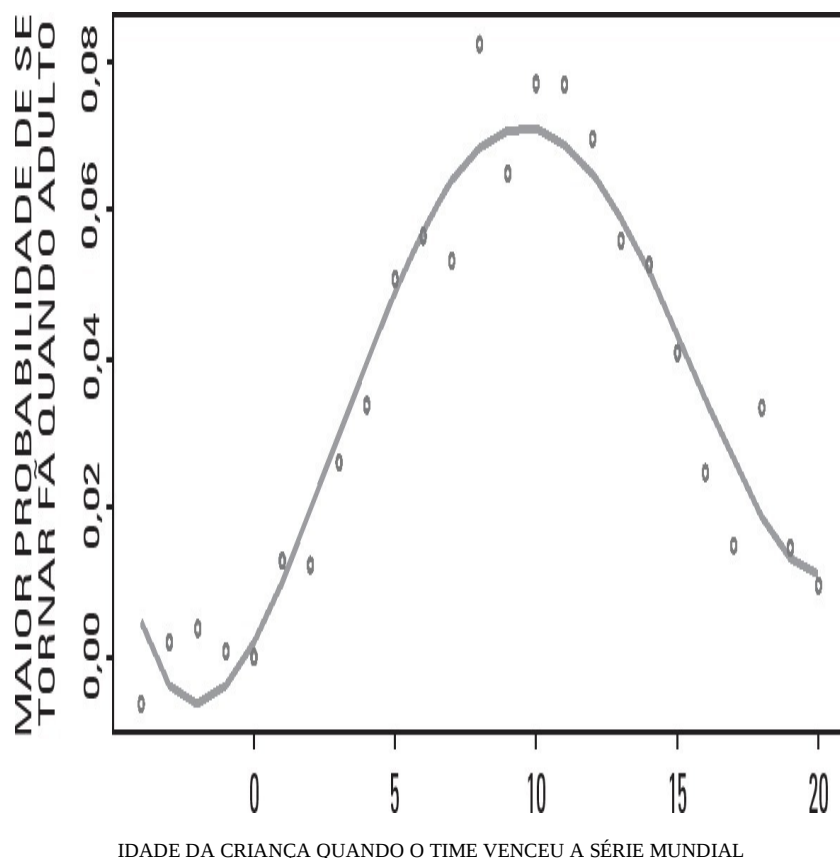
Para ver como isso funciona, vamos analisar um estudo que realizei sobre como as experiências da infância influenciam o time de beisebol para o qual você vai torcer — ou se vai ou não torcer para um time. Para este estudo, usei os dados do Facebook sobre "curtidas" de times de beisebol. (No capítulo anterior comentei que os dados do Facebook são profundamente traiçoeiros em temas delicados. Com esse estudo, estou presumindo que ninguém, nem mesmo um torcedor do Phillies, tenha vergonha de admitir no Facebook que torce para um time específico.)

Para começar, baixei os dados do número de usuários do sexo masculino de todas as idades que "curtiram" cada um dos dois times de beisebol de Nova York. Veja a seguir o porcentual que é fã do Mets, por ano de nascimento.

PERCENTUAL DE FÃS DO SEXO MASCULINO DOS TIMES DE BEISEBOL DE NOVA YORK QUE "CURTEM" OS METS, POR ANO DE NASCIMENTO

Quanto mais alto o ponto no gráfico, mais fãs do Mets. A popularidade do time aumenta e diminui, depois aumenta e diminui novamente, sendo o Mets muito popular entre aqueles nascidos em 1962 e 1978. Aposto que os fãs de beisebol têm uma boa ideia do que isso representa. O Mets venceu apenas duas Séries Mundiais: em 1969 e em 1986. Esses homens tinham aproximadamente 7 ou 8 anos quando o Mets venceu. Assim, um excelente preditor para um torcedor do Mets, pelo menos para meninos, é o Mets ter vencido uma Série Mundial quando tinham 7 ou 8 anos.

Na verdade, podemos estender esta análise. Baixei informações do Facebook mostrando quantos fãs de todas as idades “curtem” cada um dos times de uma ampla seleção de times da Liga Principal de Beisebol.

Descobri que há também um número excepcionalmente alto de fãs do sexo masculino do Baltimore Orioles nascidos em 1962 e do Pittsburgh Pirates nascidos em 1963. Estes homens eram garotos de 8 anos quando esses times foram campeões. Na verdade, calcular a idade dos picos no número de fãs para todos os times que estudei, e depois descobrir quantos anos esses fãs teriam, resultou no seguinte gráfico:

IDADE DA CRIANÇA QUANDO O TIME VENCEU A SÉRIE MUNDIAL

Mais uma vez, vemos que o ano mais importante na vida de um homem, para efeito de sedimentar sua preferência por determinado time de beisebol quando adulto, é aquele em que ele tem mais ou menos 8 anos. Em geral, o período-chave para conquistar um garoto é entre os 5 e os 15 anos. Vencer quando o rapaz já tem 19 ou 20 anos tem cerca de 1/8 da importância na determinação do time para qual irá torcer, se comparado a vencer quando ele tem 8 anos. Nesta fase, o garoto já tem um time do coração ou não terá mais.

Você pode se perguntar, e quanto às mulheres fãs de beisebol? Os padrões são muito menos distintos, mas o pico de idade parece ser aos 22 anos.

Este é meu estudo favorito. Ele relaciona dois de meus assuntos preferidos: beisebol e a origem de meus dissabores de adulto. Fui irremediavelmente fisgado em 1986 e desde então sofro — e torço — pelo Mets. Noah teve o bom senso de nascer quatro anos depois e foi poupado desse sofrimento.

Entretanto, o beisebol não é o assunto mais importante do mundo, ou pelo menos era o que me diziam incansavelmente meus orientadores de doutorado. Mas esta metodologia nos ajuda a lidar com questões semelhantes, incluindo como as pessoas desenvolvem suas preferências políticas, tendências sexuais, gosto musical e hábitos financeiros. (Eu estava particularmente interessado nas origens das ideias excêntricas de meu irmão sobre os últimos dois tópicos.) Minha predição é que descobriremos que muitos de nossos comportamentos e interesses quando adultos, mesmo aqueles que consideramos nossa essência, são explicados por fatos arbitrários de quando nascemos e pelo que estava acontecendo em certos anos-chave de nossa juventude.

Na verdade, alguns trabalhos sobre a origem das preferências políticas já foram feitos. Yair Ghitza, cientista-chefe na Catalist, uma empresa de análise de dados, e Andrew Gelman, cientista político e estatístico na Universidade de Columbia, tentaram testar a ideia convencional de que a maioria das pessoas é liberal na juventude e se torna cada vez mais conservadora à medida que envelhece. Esta é a visão expressa em uma famosa citação frequentemente atribuída a Winston Churchill: "Qualquer pessoa com menos de 30, que não seja liberal, não tem coração; e qualquer pessoa com mais de 30, que não seja conservadora, não tem cérebro."

Ghitza e Gelman examinaram sessenta anos de dados de pesquisas, utilizando mais de 300 mil observações sobre preferências eleitorais. Eles descobriram, contrariando a alegação de Churchill, que adolescentes às vezes tendem para o liberalismo e às vezes, para o conservadorismo. Assim como as pessoas de meia-idade e os idosos.

Esses pesquisadores descobriram que as visões políticas, na verdade, se formam de uma maneira muito semelhante às preferências em relação aos times. Existe um período crucial que marca a pessoa para o resto da

vida. Entre as idades críticas de 14 e 24 anos, inúmeros norte-americanos formam suas visões baseados na popularidade do presidente do momento. Um Republicano popular ou um Democrata impopular influenciam muitos jovens adultos a se tornar Republicanos. Um Republicano odiado ou um Democrata idolatrado colocam este grupo impressionável na coluna dos Democratas.

E essas visões, nestes anos críticos, durarão, na média, a vida toda.

Para ver como isso funciona, compare os norte-americanos nascidos em 1941 com os nascidos uma década mais tarde.

Aqueles no primeiro grupo chegaram à fase adulta durante a presidência de Dwight D. Eisenhower, um popular Republicano. No início da década de 1960, apesar de estar abaixo dos 30 anos, essa geração tendeu fortemente para o Partido Republicano. E seus membros continuaram com tendências para o Partido Republicano ao envelhecer.

Norte-americanos nascidos dez anos mais tarde — os *baby boomers* [pessoas nascidas entre 1946 e 1964] — atingiram a fase adulta durante a presidência de John F. Kennedy, um Democrata extremamente popular; Lyndon B. Johnson, um Democrata inicialmente adorado; e Richard M. Nixon, um Republicano que ocasionalmente renunciou em total desonra. Membros dessa geração tenderam ao liberalismo durante toda a vida.

Com todos esses dados, os pesquisadores foram capazes de determinar um único ano crucial para o desenvolvimento das visões políticas: os 18 anos.

E eles descobriram que os efeitos dessas impressões são substanciais. O modelo deles estima que a experiência de Eisenhower resultou em um aumento de cerca de 10 pontos percentuais ao longo da vida para os Republicanos entre os nascidos em 1941. A experiência de Kennedy, Johnson e Nixon forneceu aos Democratas uma vantagem de 7 pontos percentuais entre os norte-americanos nascidos em 1952.

Deixei bastante claro que sou cético em relação a dados de pesquisa, mas estou impressionado com o grande número de respostas examinadas aqui. Na verdade, este estudo não poderia ter sido feito com uma pesquisa pequena. Os pesquisadores precisaram de centenas de milhares de observações, reunidas a partir de muitas pesquisas, para ver como as preferências mudam conforme as pessoas envelhecem.

O tamanho dos dados foi crucial também para meu estudo sobre o beisebol. Precisei focar não apenas os fãs de cada time, mas as pessoas de cada idade. Milhões de observações são necessárias para fazer isso, e o Facebook e outras fontes digitais rotineiramente oferecem estes números.

É aqui que a grandiosidade do Big Data realmente entra em cena. Você precisa de muitos píxeis em uma foto para ser capaz de a ampliar com nitidez em determinado ponto. Da mesma forma, são necessárias muitas observações em um conjunto de dados para possibilitar a identificação clara de um pequeno subconjunto daqueles dados — por exemplo, qual é a popularidade do Mets entre homens nascidos em 1978. Uma pequena pesquisa de alguns milhares de pessoas não terá uma amostra grande o suficiente desses homens.

Este é o terceiro poder do Big Data: permite uma diferenciação significativa em pequenos segmentos de um conjunto de dados para obter novas perceptivas sobre quem somos. E podemos analisar de perto outras dimensões além da idade. Se tivermos dados suficientes, podemos ver como as pessoas em determinadas cidades e bairros se comportam. E podemos descobrir como as pessoas se comportam mal hora a hora ou até minuto a minuto.

Neste capítulo, o comportamento humano é o destaque.

O QUE REALMENTE ESTÁ ACONTECENDO EM NOSSOS CONDADOS, CIDADES E BAIRROS?

Em retrospecto, é surpreendente. Mas quando Raj Chetty, então professor em Harvard, e uma pequena equipe de pesquisa obtiveram pela primeira vez um conjunto de dados especialmente grande — os registros de impostos de todos os norte-americanos desde 1996 —, não tinham certeza do que conseguiriam extrair. A Receita Federal dos Estados Unidos forneceu os dados por achar que os pesquisadores pudessem ajudá-la a esclarecer os efeitos da política de tributação.

As primeiras tentativas de Chetty e de sua equipe na utilização do Big Data levaram, de fato, a inúmeros becos sem saída. As investigações das consequências das políticas de tributação estaduais e federais chegaram praticamente às mesmas conclusões que aqueles que utilizavam pesquisas comuns. Talvez as respostas de Chetty, usando centenas de milhões de pontos de dados da Receita Federal, fossem mais precisas. Mas obter as mesmas respostas de antes, com um pouco mais de precisão, não é um grande feito para a ciência social. Não é o tipo de trabalho que os principais jornais estão ávidos para publicar.

Além disso, organizar e analisar todos os dados da Receita Federal consumia muito tempo. Chetty e sua equipe — soterrados pelos dados — levavam mais tempo para obter as mesmas respostas.

Começava a parecer que os céticos do Big Data estavam certos. Não era preciso dados de centenas de milhões de norte-americanos para entender a política tributária; uma pesquisa com dez mil pessoas era o bastante. Chetty e

sua equipe ficaram bastante desanimados.

E então, finalmente, os pesquisadores perceberam seu erro. “Big Data não quer dizer fazer apenas a mesma coisa que você faria com a pesquisa, exceto pela utilização de mais dados”, explica Chetty. Eles estavam fazendo perguntas apropriadas para poucos dados para a gigantesca coleção de dados que tinham em mãos. “Big Data deveria na verdade permitir a utilização de arranjos totalmente diferentes daqueles usados na pesquisa comum”, acrescenta Chetty. “Você pode, por exemplo, focar geografias.”

Em outras palavras, com dados sobre centenas de milhões de pessoas, Chetty e sua equipe conseguiriam identificar padrões em cidades e bairros, grandes e pequenos.

Quando era aluno de pós-graduação em Harvard, fui a um seminário em que Chetty apresentou seus resultados iniciais usando os registros dos impostos de todos os norte-americanos. Cientistas sociais referem-se em seus trabalhos a observações — significando quantos pontos de dados possuem. Se um cientista social estiver trabalhando com uma pesquisa de 800 pessoas, dirá: “Temos 800 observações.” Se estivesse trabalhando em um experimento de laboratório com 70 pessoas, diria: “Temos 70 observações.”

“Temos 1,2 bilhão de observações”, disse Chetty, impassível. A plateia riu nervosamente.

Chetty e seus coautores começaram, naquela sala de conferências e então em uma série de artigos, a nos dar importantes novas percepções sobre como os Estados Unidos funcionam.

Pense na seguinte questão: os Estados Unidos são a terra da oportunidade? Você tem uma chance real, se seus pais não forem ricos, de enriquecer?

A forma tradicional de responder a essa pergunta é examinar uma amostra representativa dos norte-americanos e comparar o resultado com os dados semelhantes de outros países.

Estes são os dados de uma variedade de países sobre a igualdade de oportunidades. A pergunta foi: Qual é a chance de uma pessoa com pais entre os 20% das pessoas na faixa de menor renda atingir a faixa dos 20% de maior renda?

CHANCES DE UMA PESSOA COM PAIS POBRES SE TORNAR RICA (SELEÇÃO DE PAÍSES)	
Estados Unidos	7,5
Reino Unido	9,0
Dinamarca	11,7
Canadá	13,5

Como você pode ver, os Estados Unidos *não* se classificaram bem.

Mas essa análise simples ignora a verdadeira história. A equipe de Chetty selecionou dados com base na geografia. Eles descobriram que as probabilidades divergem muito dependendo do lugar dos Estados Unidos em que a pessoa nasceu.

CHANCES DE UMA PESSOA COM PAIS POBRES SE TORNAR RICA (SELEÇÃO DE PARTES DOS ESTADOS UNIDOS)	
San Jose, CA	12,9
Washington, DC	10,5
Média dos Estados Unidos	7,5
Chicago, IL	6,5
Charlotte, NC	4,4

Em algumas partes dos Estados Unidos, a probabilidade de uma criança nascida pobre ter sucesso é igual à de qualquer país desenvolvido no mundo. Em outras partes dos Estados Unidos, a probabilidade de uma criança pobre ter sucesso é menor do que em qualquer outro país.

Esses padrões nunca seriam observados em uma pesquisa pequena, que pode apenas não incluir algumas pessoas em Charlotte e San Jose, e que, portanto, não permitiria a distinção.

Na verdade, a equipe de Chetty poderia focar ainda mais. Como tinham uma enorme quantidade de dados — sobre todos os norte-americanos —, eles conseguiriam focar pequenos grupos de pessoas que se mudavam de cidade em cidade para ver como isto afetaria suas chances de sucesso: aqueles que mudaram da cidade de Nova York para Los Angeles, de Milwaukee para Atlanta, de San Jose para Charlotte. Isto permitiu a eles testarem para identificar a causa, não apenas a correlação (uma distinção que discuto no próximo capítulo). E, sim, mudar-se para a cidade certa nos anos de formação fez uma diferença significativa.

Então, os Estados Unidos são "a terra da oportunidade"?

A resposta não é sim nem não. A resposta é: algumas partes, sim e outras, não.

Quando autores escrevem: "Os Estados Unidos são melhor descritos como uma coleção de sociedades, algumas delas são 'terras de oportunidade' com altas taxas de mobilidade ao longo de gerações, e há outras em que poucas crianças escapam da pobreza."

Então qual é o problema nas partes dos Estados Unidos em que há alta taxa de mobilidade de renda? O que torna alguns lugares melhores em igualar as condições, em permitir que uma criança pobre tenha uma vida muito boa? Áreas que gastam mais em educação oferecem uma melhor oportunidade para as crianças pobres. Lugares com mais pessoas religiosas e menor criminalidade se saem melhor. Locais com mais pessoas negras se saem pior. Um fato interessante é que isso teve um efeito não apenas sobre as crianças negras, mas também sobre as brancas vivendo na mesma localidade. Áreas com muitas mães solteiras se saem pior. Este efeito é o mesmo não apenas para os filhos de mães solteiras, mas para os de pais casados morando em locais com muitas mães solteiras. Alguns desses resultados sugerem que os amigos de uma criança pobre fazem diferença. Se seus amigos vêm de um contexto social difícil e com poucas oportunidades, ela tem que se esforçar mais para escapar da pobreza.

Os dados nos dizem que algumas partes dos Estados Unidos são melhores em oferecer oportunidades para as crianças escaparem da pobreza. Então, quais lugares proporcionam maiores chances para as pessoas escaparem do anjo da morte?

Gostamos de pensar na morte como um grande equalizador. Ninguém, afinal, é capaz de a evitar. Nem o miserável nem o rei, nem o indigente nem Mark Zuckerberg. Todo mundo morre.

Mas se os ricos não podem escapar da morte, os dados nos mostram que, hoje, eles conseguem adiá-la. Mulheres norte-americanas pertencentes ao 1% de maior renda vivem, em média, dez anos a mais do que as norte-americanas pertencentes ao 1% de menor renda. Para os homens, a lacuna é de quinze anos.

Como esses padrões variam em diferentes partes dos Estados Unidos? A expectativa de vida se altera conforme o local em que você mora? Esta variável é diferente para pessoas ricas e pobres? De novo, ao focarem uma geografia, a equipe de Raj Chetty encontrou as respostas.

Um fato interessante, para os norte-americanos mais abastados, é que a expectativa de vida quase não é afetada pelo local em que moram. Se você tem muito dinheiro, pode esperar alcançar aproximadamente os 89 anos sendo mulher e os 87 sendo homem. Os ricos de todos os lugares tendem a cultivar hábitos de saúde mais saudáveis — em média, se exercitam mais, se alimentam melhor, fumam menos e têm menor probabilidade de sofrer de obesidade. Eles podem bancar a esteira, os abacates orgânicos e as aulas de yoga. E conseguem adquirir estas coisas em qualquer recanto dos Estados Unidos.

Para os pobres, a história é bastante diferente. Para eles, a expectativa de vida varia enormemente dependendo de onde moram. Na verdade, viver no lugar certo pode acrescentar cinco anos à expectativa de vida de uma pessoa pobre.

Então, por que alguns lugares parecem permitir aos depauperados viver mais? Quais atributos as cidades em que as pessoas pobres vivem mais têm em comum?

Aqui estão quatro dos atributos de uma cidade — três deles não se correlacionam com a expectativa de vida dos mais pobres, e um deles, sim. Veja se consegue adivinhar qual.

O QUE FAZ AS PESSOAS POBRES DE UMA CIDADE VIVEREM MAIS?

Alto nível de religiosidade.
Baixos níveis de poluição.
Um porcentual mais alto de residentes cobertos por planos de saúde.
Muitos ricos morarem na cidade.

As primeiras três — religião, ambiente e plano de saúde — não se correlacionam com expectativas de vida mais elevadas para os pobres. Qual é a variável relevante, de acordo com Chetty e outros que trabalharam neste estudo? Quantas pessoas ricas moravam na cidade. Mais ricos em uma cidade significam que os residentes pobres

têm vidas mais longas. Pessoas pobres na cidade de Nova York, por exemplo, vivem muito mais do que os pobres em Detroit.

Por que a presença de ricos é um preditor tão poderoso para a expectativa de vida das pessoas mais pobres? Uma hipótese — e é mera especulação — foi proposta por David Cutler, um dos autores do estudo e um dos meus consultores. Comportamento contagiante pode impulsionar esse efeito.

Há um grande número de pesquisas mostrando que os hábitos são contagiosos. Então, pessoas pobres que moram perto de ricas adquirem muitos de seus hábitos. Alguns desses hábitos — digamos, um vocabulário pretensioso — não têm probabilidade de afetar a saúde de uma pessoa. Outros — como se exercitar — definitivamente terão um impacto positivo. De fato, pessoas pobres vivendo perto de ricas se exercitam mais, fumam menos e têm menor probabilidade de desenvolver obesidade.

Meu estudo favorito da equipe de Raj Chetty, que obteve acesso a uma gigantesca coleção de dados da Receita Federal, foi sua investigação sobre por que algumas pessoas mentem em seus impostos enquanto outras, não. Explicar este estudo é um pouco mais complicado.

A chave é saber que existe uma maneira fácil de pessoas autônomas com um filho maximizarem o dinheiro que recebem do governo. Se você declarar que obteve renda tributável de exatamente US$9 mil em determinado ano, o governo lhe restituirá o valor de US$1.377 — este montante representa o crédito tributário sobre a renda auferida, uma concessão para suplementar os rendimentos do trabalhador pobre, menos os impostos incidentes sobre a folha de pagamento. Declarar mais que isso faz com que seu imposto de renda aumente. Declarar menos que isso, com que o crédito tributário sobre renda auferida diminua. Uma renda tributável de US$9 mil é o número mágico.

E, você não teria como saber, US$9 mil é a renda tributada mais comum declarada por trabalhadores autônomos com um filho.

Esses norte-americanos simplesmente ajustam seus cronogramas de trabalho para ter certeza de que sua renda totalize exatamente este valor? Nada disto. Quando esses trabalhadores foram aleatoriamente auditados — uma ocorrência muito rara — quase sempre descobriu-se que seus rendimentos não foram sequer próximos de US$9 mil — receberam substancialmente menos ou mais.

Em outras palavras, eles mentiram em suas declarações de impostos fingindo ganhar o valor que lhes daria a restituição mais gorda do governo.

Assim, quão típico é esse tipo de fraude e quem dentre os profissionais autônomos com um filho tem mais probabilidade de a cometer? Ocorre que, Chetty e seus pares relataram gigantescas diferenças dentro dos Estados Unidos na frequência desse tipo de fraude. Em Miami, entre as pessoas desta categoria, impressionantes 30% declararam renda de US$9 mil. Na Filadélfia, apenas 2%.

O que prediz quem vai mentir? O que há nesses lugares com maiores números de fraudadores e naqueles com números menores? Conseguimos correlacionar os índices de fraude com outros fatores demográficos da cidade e descobrimos dois fortes preditores: uma alta concentração de pessoas na área se qualificando para o crédito tributário sobre renda auferida e uma alta concentração de profissionais de contabilidade nos arredores.

O que estes fatores indicam? Chetty e os autores tinham uma explicação. O principal motivador para fraudar seus impostos dessa maneira era a informação.

A maioria dos contribuintes autônomos com um filho simplesmente não sabia que o número mágico para receber um gordo reembolso do governo era US$9 mil. Mas morar perto de quem sabe — sejam seus vizinhos ou contadores — aumentou drasticamente as chances de virem a conhecer este fato.

Na verdade, a equipe de Chetty descobriu ainda mais evidências de que o conhecimento impulsionou este tipo de fraude. Quando os norte-americanos se mudaram de uma área em que essa variedade de fraude tributária era baixa para uma em que era alta, descobriram e adotaram a trapaça. Ao longo do tempo, a fraude se espalhou de região para região por todo os Estados Unidos. Assim como um vírus, a fraude nos impostos é contagiosa.

Agora, pare por um momento e pense sobre a relevância deste estudo. Ele demonstrou que, quando se trata de descobrir quem vai fraudar seus impostos, a chave não é determinar quem é honesto ou desonesto. É determinar quem sabe ou não como o fazer.

Então, quando alguém diz a você que jamais fraudaria sua declaração de impostos, há uma chance muito grande de estar — você adivinhou — mentindo. A pesquisa de Chetty sugere que muitas pessoas fraudariam se soubessem como.

Se quiser mentir em sua declaração de impostos (e *não* estou recomendando que o faça), você deve morar perto de profissionais de contabilidade ou perto de fraudadores de impostos que possam lhe mostrar o caminho.

Mas se quiser ter filhos mundialmente famosos, onde deve morar? Esta capacidade de focar dados e obter real especificidade pode responder essa pergunta, também.

Fiquei curioso em saber de onde vinham os norte-americanos de maior sucesso, então um dia decidi baixar os dados da Wikipédia. (Hoje em dia é possível fazer isso.)

Com um pouco de codificação, obtive um conjunto de dados de mais de 150 mil norte-americanos considerados pelos editores da Wikipédia como sendo notáveis o bastante para merecer uma entrada. O conjunto de dados incluiu os condados e as datas de nascimento, a ocupação e o gênero. Juntei estes dados com os de nascimento por condado coletados pelo Centro Nacional de Estatísticas de Saúde. Calculei as chances de aparecer na Wikipédia para as pessoas nascidas em cada condado dos Estados Unidos.

Ter seu perfil inserido na Wikipédia é um marcador significativo de feitos notáveis? Existem certas limitações. Os editores da Wikipédia tendem a ser homens jovens, o que pode distorcer a amostra. E alguns tipos de notoriedade não são particularmente honrosas. Ted Bundy, por exemplo, tem entrada na Wikipédia por ter matado dezenas de jovens mulheres. Entretanto, fui capaz de remover os criminosos da amostra sem afetar muito os resultados.

Limitei os estudos à geração *baby boomer*, pois eles tiveram praticamente uma vida inteira para se tornar notáveis.

Aproximadamente 1 a cada 2.058 norte-americanos nascidos na geração *baby boomer* foi considerado notável o suficiente para merecer uma entrada na Wikipédia. Cerca de 30% conquistou a menção através de feitos na área de arte ou entretenimento, 29% através de esportes, 9% por meio da política e 3% através do meio científico ou acadêmico.

O primeiro fato impressionante que observei nos dados foi a enorme variação geográfica na probabilidade de se tornar um grande sucesso, pelo menos em termos de Wikipédia. As chances de alcançar notoriedade eram altamente dependentes de onde nasceram. Aproximadamente 1 a cada 1.209 *baby boomers* nascidos na Califórnia foi citado na Wikipédia. Mas apenas 1 a cada 4.496 *baby boomers* nascidos na Virgínia Ocidental conseguiu o feito. Ao selecionarmos por condado, os resultados se tornam ainda mais reveladores. Aproximadamente 1 a cada 748 *baby boomers* nascidos no Condado de Suffolk, Massachusetts, em que Boston se localiza, recebeu entradas na Wikipédia. Em alguns outros condados, a taxa de sucesso foi vinte vezes menor.

Por que algumas partes do país parecem ser eficazes em produzir norte-americanos notáveis? Examinei cuidadosamente os principais condados. Os resultados mostraram que praticamente todos eles se encaixam em uma das duas categorias.

Primeiro, e isso me surpreendeu, muitos desses condados continham uma cidade universitária de tamanho considerável. Quase todas as vezes que vi um nome de condado que não havia identificado na lista de condados melhores classificados, como Washtenaw, Michigan, descobri que continha uma cidade universitária tradicional, neste caso Ann Arbor. Os condados agraciados por Madison, Wisconsin; Athens, Georgia; Columbia, Missouri; Berkeley, California; Chapel Hill, Carolina do Norte; Gainesville, Flórida; Lexington, Kentucky e Ithaca, Nova York, estão todos entre os 3% com maiores ocorrências.

Por quê? Parte pode ser explicada pela genética: filhos e filhas de professores e alunos graduados tendem a ser inteligentes (uma característica que, no jogo do sucesso, é muito útil). E, de fato, ter mais pessoas com formação universitária em uma área é um forte preditor do sucesso das pessoas nascidas naquela região.

Mas ao que tudo indica tem algo a mais por trás deste resultado: a exposição precoce à inovação. Um dos campos em que as cidades universitárias são mais bem-sucedidas em produzir líderes é a música. Uma criança em uma cidade universitária será exposta a shows únicos, estações de rádio incomuns e até lojas de gravações independentes. E isso não se limita às artes. Cidades universitárias também formam mais do que a parcela normal de pessoas de negócios notáveis. Pode ser que a exposição precoce à arte e ideias inovadoras ajude-os também.

O sucesso das cidades universitárias não apenas ignora regiões. Ignora raças. Norte-americanos são notavelmente sub-representados na Wikipédia em áreas não esportivas, especialmente em negócios e pesquisa. Isto, sem dúvida, tem muito a ver com discriminação. Mas um pequeno condado, em que a população em 1950 era 84% negra, produziu *baby boomers* relevantes em uma taxa parecida com aquelas dos contados melhor classificados.

Dentre os menos de 13 mil *baby boomers* nascidos no Condado de Macon, Alabama, 15 chegaram à Wikipédia — ou 1 a cada 852. Todos são negros. Catorze deles eram da cidade de Tuskegee, lar da Universidade de Tuskegee, uma universidade historicamente negra fundada por Booker T. Washington. A lista inclui juízes, escritores e cientistas. Na verdade, uma criança negra nascida em Tuskegee tinha a mesma probabilidade de se tornar famosa em

um campo fora do esporte do que uma branca nascida em uma das cidades universitárias de maioria branca mais bem classificadas.

O segundo atributo mais provável para formar pessoas de sucesso em um respectivo condado foi a presença no condado de uma grande cidade. Nascer no Condado de San Francisco, no Condado de Los Angeles ou na cidade de Nova York, ofereceu as probabilidades mais altas de se chegar à Wikipédia. (Agrupei os cinco condados da cidade de Nova York, pois muitas entradas da Wikipédia não especificam o distrito de nascimento.)

As áreas urbanas tendem a ser bem abastecidas com modelos de sucesso. Para analisar o valor de estar perto de profissionais de sucesso de determinada área quando jovem, compare as cidades de Nova York, Boston e Los Angeles. Dentre as três, Nova York tem as taxas mais altas de jornalistas notáveis; Boston tem as mais altas de cientistas relevantes e Los Angeles produz atores renomados em taxas mais altas. Lembre-se, estamos falando de pessoas nascidas no local, não que se mudaram para lá. E isto permanece verdadeiro até mesmo depois de subtrairmos as pessoas com pais notáveis na área.

Condados em áreas suburbanas, a menos que contenham cidades universitárias importantes, tiveram um desempenho muito pior do que suas contrapartes urbanas.

Meus pais, como muitos *baby boomers*, mudaram-se de uma região urbana supermovimentada para uma mais tranquila — neste caso, de Manhattan para o Condado de Bergen, Nova Jersey — para criar os três filhos. Isto foi potencialmente um erro, pelo menos de uma perspectiva de criar filhos célebres. Uma criança nascida na cidade de Nova York tem 80% mais probabilidade de aparecer na Wikipédia do que uma criança nascida no Condado de Bergen. Essas são apenas correlações, mas sugerem que crescer perto das grandes ideias é melhor do que com um grande quintal.

Os severos efeitos identificados aqui podem ser ainda mais fortes se eu tivesse melhores dados sobre os lugares em que viveram durante a infância, já que muitas pessoas cresceram em condados diferentes daqueles em que nasceram.

O sucesso das cidades universitárias e das grandes cidades é surpreendente quando apenas olhamos os dados. Mas eu também me aprofundei mais para fazer uma análise empírica mais sofisticada.

Isto mostrou que havia outra variável que foi um forte preditor de presença de uma pessoa na Wikipédia: a proporção de imigrantes em seus condados de nascimento. Quanto maior o porcentual de residentes estrangeiros, maior a proporção de crianças nascidas naquela área que obterão um sucesso representativo. (Engula isto, Donald Trump!) Se dois lugares têm populações urbanas e universitárias parecidas, aquela com mais imigrantes produzirá mais norte-americanos proeminentes. O que explica isto?

Em grande parte, isto parece ser diretamente atribuível aos filhos de imigrantes. Fiz uma busca exaustiva nas biografias dos cem *baby boomers* brancos mais famosos, de acordo com o projeto Pantheon, do Instituto de Tecnologia de Massachusetts (MIT), que também trabalha com os dados da Wikipédia. A maioria deles era do setor de entretenimento. Pelo menos treze tinham mães estrangeiras, incluindo Oliver Stone, Sandra Bullock e Julianne Moore. Esta taxa é mais de três vezes maior do que a média nacional durante o período. (Muitos tinham pais imigrantes, incluindo Steve Jobs e John Belushi, mas esses dados eram mais difíceis de comparar com as médias nacionais, pois nem sempre as informações sobre os pais constam das certidões de nascimento.)

E quanto às variáveis que não impactam no sucesso? Uma que achei um pouco surpreendente foi quanto o estado gasta em educação. Em estados com porcentuais semelhantes de seus residentes vivendo em áreas urbanas, os gastos com educação não se correlacionaram com as taxas de produção de escritores célebres, artistas ou líderes de negócios.

É interessante comparar meu estudo da Wikipédia com aquele realizado pela equipe de Chetty, discutido anteriormente. A equipe de Chetty tentou descobrir quais áreas são boas em permitir que as pessoas alcancem a classe média alta. Meu estudo tentava desvendar quais áreas são boas em permitir que as pessoas alcancem a fama. Os resultados são visivelmente diferentes.

Altos gastos em educação ajudam as crianças a chegar à classe média alta. Mas não ajudam muito em as tornar escritoras, artistas ou líderes de negócios representativos, pois muitas dessas pessoas de enorme sucesso odiavam a escola e algumas chegaram a abandoná-la.

A cidade de Nova York, descobriu a equipe de Chetty, não é um local particularmente bom para criar uma criança se pretende garantir que ela chegue à classe média alta. É um ótimo lugar, meu estudo revelou, se quiser dar a seu filho uma chance de fama.

Quando se analisa os fatores que impulsionam o sucesso, uma grande variação entre os condados começa a fazer sentido. Muitos reúnem os principais ingredientes para o sucesso. Voltemos, mais uma vez, para Boston. Com

inúmeras universidades, Boston fervilha com ideias inovadoras. É uma área urbana com muitas pessoas extremamente bem-sucedidas oferecendo aos jovens exemplos de como chegar lá. E atrai numerosos imigrantes, cujos filhos são incentivados a colocar essas lições em prática.

E se uma área não tiver quaisquer dessas qualidades? Está condenada a produzir menos superestrelas? Não necessariamente. Existe outro caminho: especialização extrema. O Condado de Roseau, Minnesota, um pequeno condado rural com poucos estrangeiros e nenhuma universidade importante, é um bom exemplo. Aproximadamente 1 a cada 740 pessoas nascidas aqui conquistou uma vaga na Wikipédia. O segredo? Todos os nove eram jogadores de hóquei profissional, sem dúvida ajudados pelos excepcionais programas de hóquei para jovens e estudantes do ensino médio.

Então, a questão é — presumindo que você não esteja interessado em criar uma estrela do hóquei — mudar para Boston ou Tuskegee se quiser dar a seus futuros filhos a melhor vantagem possível? Mal não faz. Mas há uma lição maior aqui. Geralmente, economistas e sociólogos se concentram em evitar resultados ruins, como pobreza e criminalidade. Ainda assim, o objetivo de uma sociedade excepcional não é apenas deixar menos pessoas para trás; é ajudar o maior número possível a realmente se sobressair. Talvez este esforço para identificar os lugares em que centenas de milhares de norte-americanos famosos nasceram possa nos oferecer algumas estratégias iniciais: estimular a imigração, subsidiar universidades e apoiar as artes, entre eles.

Normalmente, estudo os Estados Unidos. Então, quando quero focar a geografia, penso em nossas cidades e bairros — em analisar lugares como o Condado de Macon, Alabama, e o de Roseau, Minnesota. Mas outra vantagem gigantesca — e ainda crescente — dos dados da internet é que são facilmente coletados de todo o mundo. Podemos então ver como os países diferem. E os cientistas de dados têm uma oportunidade de explorar a antropologia.

Um assunto um tanto aleatório que explorei recentemente: como é a gravidez em diferentes países ao redor do mundo? Examinei as buscas no Google feitas por grávidas. A primeira coisa que descobri foi a impressionante similaridade nos sintomas dos quais as mulheres reclamam.

Testei com que frequência diversos sintomas foram pesquisados combinados com a palavra “grávida”. Por exemplo, com que frequência o termo “grávida” é pesquisado junto com “náusea”, “dor nas costas” ou “constipação”? Os sintomas no Canadá foram muito parecidos com aqueles dos Estados Unidos. Sintomas em países como Inglaterra, Austrália e Índia foram também muito similares.

Grávidas em todo o mundo aparentemente também têm os mesmos desejos. Nos Estados Unidos, a principal busca no Google nessa categoria é “desejo por gelo durante a gravidez”. As quatro seguintes são sal, doces, frutas e comidas apimentadas. Na Austrália, esses desejos não diferem tanto: a lista contém sal, doces, chocolate, gelo e fruta. E quanto à Índia? Um cenário parecido: comidas picantes, doces, chocolate, sal e sorvete. Na verdade, os cinco desejos mais comuns são parecidos em todos os países que pesquisei.

Evidências preliminares sugerem que nenhuma parte do mundo encontrou uma dieta ou ambiente que mudem drasticamente a experiência física da gravidez.

Mas os pensamentos que a cercam definitivamente apresentam diferenças.

Começando pelas perguntas sobre o que grávidas podem fazer com segurança. As principais perguntas nos Estados Unidos são: grávidas podem “comer camarão”, “beber vinho”, “tomar café” ou “tomar Tylenol”?

Quando se trata destas preocupações, outros países não têm muito em comum com os Estados Unidos ou entre si. Se uma grávida pode ou não “beber vinho” não está entre as dez questões mais comuns no Canadá, Austrália ou Inglaterra. As preocupações na Austrália são principalmente relacionadas à ingestão de laticínios, especialmente cream cheese. Na Nigéria, em que 30% da população utiliza a internet, a principal pergunta é se grávidas podem beber água gelada.

Estas preocupações são legítimas? Depende. Há fortes evidências de que grávidas têm um risco aumentado de contrair listeriose de queijo não pasteurizado. Foram estabelecidas ligações entre ingerir muito álcool e efeitos negativos para o bebê. Em algumas partes do mundo, acredita-se que beber água gelada provoca pneumonia no bebê; não conheço médico algum que endosse isto.

As enormes diferenças nas perguntas em todo o mundo são muito provavelmente causadas por uma desmedida enxurrada de informações vindas de fontes discrepantes em cada país: estudos científicos legítimos, estudos medíocres, superstições e conversas de vizinhos. É difícil para as mulheres saberem no que prestar atenção — ou o que pesquisar no Google.

Podemos ver outra clara diferença quando analisamos as buscas mais comuns na forma de: “Como… durante a gravidez?” Nos Estados Unidos, na Austrália e no Canadá, a busca mais frequente é: “Como prevenir estrias durante

a gravidez.” Mas em Gana, na Índia e na Nigéria, prevenir estrias não aparece sequer entre as cinco mais populares. Estes países tendem a se preocupar mais em como fazer sexo ou dormir durante a gravidez.

CINCO PRINCIPAIS BUSCAS (EM ORDEM) POR “COMO… DURANTE A GRAVIDEZ”

ESTADOS UNIDOS	ÍNDIA	AUSTRÁLIA	INGLATERRA	NIGÉRIA	ÁFRICA DO SUL
prevenir estrias	dormir	prevenir estrias	perder peso	fazer sexo	fazer sexo
perder peso	transar	perder peso	prevenir estrias	perder peso	perder peso
fazer sexo	fazer sexo	evitar estrias	evitar estrias	fazer amor	prevenir estrias
evitar estrias	sexo	dormir	dormir	se manter saudável	dormir
ficar em forma	se cuidar	fazer sexo	fazer sexo	parar de vomitar	parar de vomitar

CINCO PRINCIPAIS BUSCAS COMEÇANDO POR “GRÁVIDAS PODEM…?”

ESTADOS UNIDOS	comer camarão	beber vinho	tomar café	tomar Tylenol	comer sushi
INGLATERRA	comer camarão	comer salmão defumado	comer cheesecake	comer muçarela	comer maionese
AUSTRÁLIA	comer cream cheese	comer camarão	comer bacon	comer creme azedo	comer queijo feta
NIGÉRIA	beber água gelada	beber vinho	tomar café	fazer sexo	comer moringa (planta comestível)
CINGAPURA	tomar chá-verde	tomar sorvete	comer durião	tomar café	comer abacaxi
ESPANHA	comer patê	comer presunto	tomar paracetamol	comer atum	tomar sol
ALEMANHA	voar	comer salame	ir à sauna	comer mel	comer muçarela
BRASIL	pintar o cabelo	tomar Dipirona	tomar paracetamol	andar de bicicleta	voar

Sem dúvida há muito mais a se aprender sobre analisar determinados aspectos da saúde e da cultura em diferentes cantos do mundo. Mas minha análise preliminar sugere que o Big Data nos mostrará que os humanos são ainda menos poderosos do que pensamos quando se trata de transcender a biologia. Ainda assim inventamos interpretações notavelmente diferentes do que isso tudo significa.

COMO PREENCHEMOS NOSSAS HORAS E MINUTOS

“As aventuras de um jovem cujos principais interesses são estupro, violência extrema e Beethoven.”

Foi assim que o controverso *Laranja Mecânica*, de Stanley Kubrick, foi anunciado. No filme, o protagonista, Alex DeLarge, comete atos chocantes de violência com assustadora indiferença. Em uma das cenas mais famosas do filme, ele estupra uma mulher enquanto a espanca ao ritmo de “Singin’ in the Rain” [“Cantando na Chuva”, em tradução livre].

Quase imediatamente, houve relatos de incidentes de imitação. Na verdade, um grupo de homens estuprou uma garota de 17 anos cantando a mesma canção. O filme foi banido em muitos países da Europa e algumas de suas cenas mais chocantes, removidas da versão exibida nos Estados Unidos.

Existem, de fato, muitos exemplos da vida imitando a arte, com homens aparentemente hipnotizados por aquilo que haviam visto na tela. Uma exibição de filme sobre gangues, *As Cores da Violência*, foi seguida por um violento

tiroteio. Uma apresentação do filme de gangues *New Jack City: A Gangue Brutal* foi seguida de episódios de perturbação da ordem e baderna.

Talvez o fato mais desconcertante, quatro dias depois do lançamento de *Assalto sobre Trilhos*, foram homens usando fluido de esqueiro para incendiar uma cabine de bilheteria do metrô, imitando quase perfeitamente uma cena do filme. A única diferença entre o incêndio de verdade e o da ficção: no filme, o funcionário escapou. Na vida real, morreu queimado.

Ainda há alguma evidência extraída de experimentos psicológicos de que sujeitos expostos a um filme violento reportam mais raiva e hostilidade, mesmo que não imitem precisamente qualquer cena do filme.

Em outras palavras, relatos e experimentos sugerem que filmes violentos incitam comportamentos agressivos. Mas qual é o real tamanho de seu impacto? Estamos falando de um ou dois homicídios a cada década ou centenas por ano? Relatos e experimentos não conseguem responder isto.

Para ver se o Big Data é capaz deste feito, dois economistas, Gordon Dahl e Stefano DellaVigna, juntaram três conjuntos de Big Data dos anos de 1995 a 2004: dados de crimes por hora do FBI, números de bilheterias e uma avaliação de violência em cada filme do site kids-in-mind.com [conteúdo em inglês].

A informação que eles usaram era completa — todo filme e todo crime cometido em cada hora nas cidades em todos os Estados Unidos. Isto se mostraria relevante.

A chave do estudo deles eram os finais de semana em que o filme mais popular era violento — *Hannibal* ou *Despertar dos Mortos*, por exemplo — versus os outros, em que o filme não o era, como *Noiva em Fuga* ou *Toy Story*.

Os economistas puderem ver exatamente quantos homicídios, estupros e agressões foram cometidos nos finais de semana quando um filme violento de renome foi lançado e comparar estas informações aos números de homicídios, estupros e agressões ocorridos nos finais de semana em que um filme célebre pacífico saiu.

Então, o que descobriram? Quando um filme violento era exibido, os crimes aumentavam, como alguns experimentos sugerem? Ou permaneciam nos mesmos patamares?

Nos finais de semana com um filme violento popular, os economistas descobriram, os crimes diminuíram.

Sim, você leu direito. Quando o filme em voga era violento, quando milhões de norte-americanos eram expostos a imagens de homens matando outros, os crimes diminuíam — significativamente.

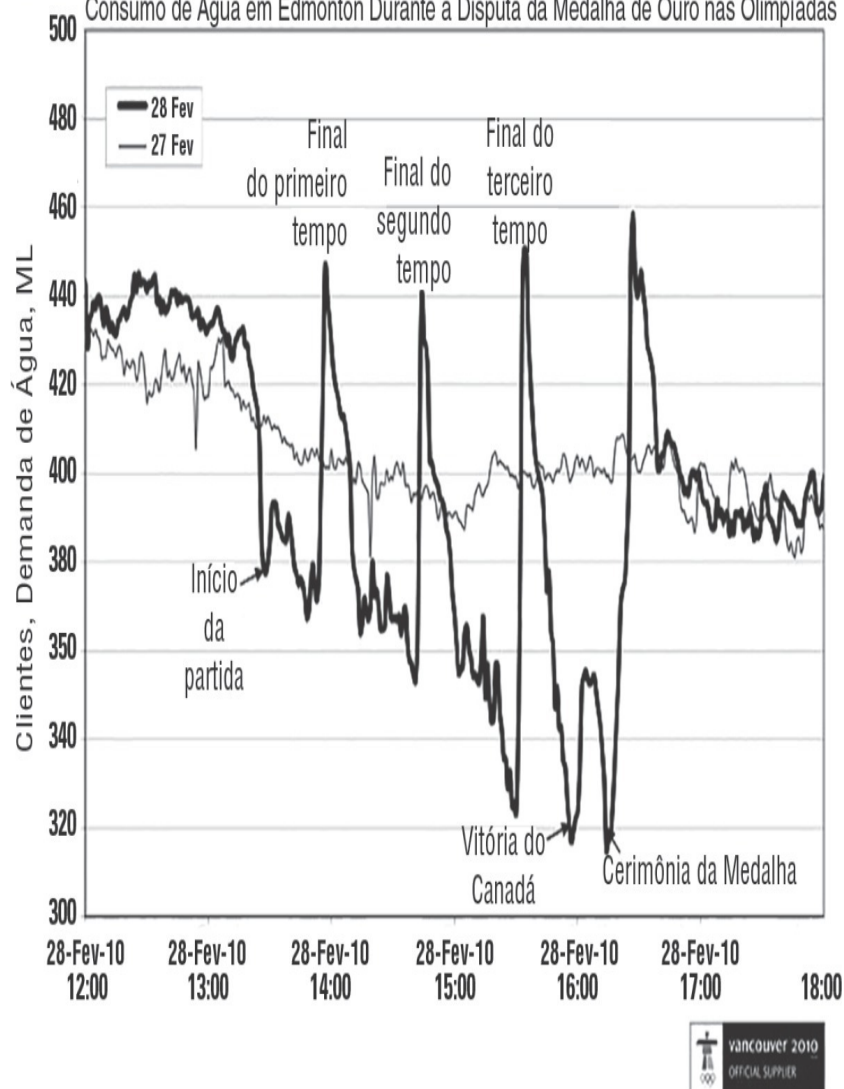
Quando obtemos um resultado estranho e inesperado como este, o primeiro pensamento é que fizemos algo errado. Cada um dos autores reanalisou cuidadosamente a compilação. Nenhum erro. O segundo pensamento é que deve haver outra variável que explique os resultados. Eles checaram se a época do ano afetava os resultados. Não afetava. Eles coletaram dados sobre o clima, pensando que talvez de algum modo isto direcionasse a relação. Nada.

“Checamos todas nossas suposições, tudo que estávamos fazendo”, conta Dahl. “Não conseguimos descobrir nada de errado.”

Apesar dos relatos e das evidências laboratoriais, por mais bizarro que possa parecer, a exibição de um filme violento de algum modo provocou uma grande queda nos crimes. Como é possível?

A chave para Dahl e DellaVigna desvendarem este mistério foi utilizar Big Data. Dados de pesquisa tradicionalmente ofereciam informação anual ou, quando muito, mensal. Se tivermos muita sorte, obteremos os dados de um final de semana. Por comparação, à medida que usamos conjuntos de dados cada vez mais abrangentes, em vez de pequenas amostras de pesquisas, nos tornamos capazes de nos concentrar em dados de uma hora e até de um minuto. Isto nos permite aprender muito mais sobre o comportamento humano.

Às vezes as flutuações ao longo do tempo são divertidas, às vezes, chocantes. A EPCOR, uma empresa de serviços em Edmonton, Canadá, relatou o consumo de água minuto a minuto durante a partida de disputa da medalha de ouro no hóquei entre Estados Unidos e Canadá, nas Olimpíadas de 2010, assistida por 80% dos canadenses, segundo as estimativas. Os dados nos mostram que logo após o fim de cada tempo, o consumo de água aumentava muito. Os banheiros em toda Edmonton disparavam suas descargas.

As buscas no Google também são discriminadas minuto a minuto, revelando alguns padrões interessantes no processo. Por exemplo, as buscas por “jogos desbloqueados” explodem às 8h da manhã em dias de semana e permanecem altas até às 15h, sem dúvida em resposta às tentativas das escolas de bloquear o acesso aos jogos nos celulares sem bani-los dos alunos.

As taxas de busca por "clima", "oração" e "notícias" têm seu pico antes das 5h30 da manhã, prova de que a maioria das pessoas acorda muito mais cedo do que eu. As taxas de busca por "suicídio" alcançam seu ápice às 00h36, e são mais baixas por volta das 9h, evidência de que a maioria das pessoas se sente bem menos infeliz de manhã do que eu.

Os dados mostram que entre as 2h e as 4h da manhã e o horário nobre são para os grandes questionamentos: "Qual é o significado de consciência? Existe livre-arbítrio? Há vida em outros planetas?" A popularidades destas perguntas à noite pode ser resultado, em parte, do consumo de maconha. As taxas de busca para "como enrolar um baseado" têm seu pico entre 1h e 2h da manhã.

E neste amplo conjunto de dados, Dahl e DellaVigna poderiam analisar como os crimes mudavam hora a hora nos finais de semana daqueles filmes. Eles descobriram que a queda nos crimes quando filmes violentos populares eram exibidos — em comparação a outros finais de semana — começam no início da noite. A criminalidade foi mais baixa, em outras palavras, antes que as cenas de violência sequer tivessem começado, quando os frequentadores do cinema ainda estavam a caminho.

Você consegue adivinhar o porquê? Pense, primeiro, sobre quem é propenso a assistir a um filme violento. São homens — especialmente jovens e agressivos.

Pense, então, sobre onde os crimes tendem a ser cometidos. Raramente em uma sala de cinema. Houve exceções, a mais famosa o caso do tiroteio premeditado em um cinema no Colorado, em 2012. Mas, em geral,

homens vão ao cinema desarmados e se sentam tranquilamente.

Ofereça uma oportunidade para homens jovens agressivos assistir *Hannibal*, e eles irão correndo ao cinema. Diante da possibilidade de assistir a *Noiva em Fuga*, homens jovens e agressivos preferem, talvez, ir a um bar, a um clube noturno ou a um salão de sinuca, onde a incidência de crimes é maior.

Filmes violentos, potencialmente, tiram as pessoas agressivas das ruas.

Mistério resolvido. Certo? Ainda não. Havia mais uma coisa estranha nos dados. Os efeitos começavam exatamente no início da exibição; entretanto, não paravam depois do término dos filmes e do fechamento do cinema. Nas noites em que os filmes violentos foram exibidos, a criminalidade diminuiu, da meia-noite às 6h.

Mesmo que ela tenha sido menor enquanto os jovens estavam no cinema, ela não deveria aumentar depois que eles saíssem e não estivessem mais ocupados? Eles acabaram de assistir a um filme violento, o que os experimentos dizem que torna as pessoas mais irritadas e agressivas.

Você consegue imaginar quaisquer explicações para o porquê de os crimes ainda caírem depois que o filme acabou? Depois de muito pensar, os autores, especialistas em criminalidade, tiveram outro momento de revelação. Eles sabiam que o álcool é um grande contribuinte para o crime. Os autores haviam frequentado cinemas suficientes para saber que praticamente nenhum nos Estados Unidos serve bebida alcoólica. Na verdade, os autores descobriram que os crimes relacionados ao consumo de álcool caíram vertiginosamente nas horas após a exibição dos filmes violentos.

Obviamente, os resultados de Dahl e DellaVigna eram limitados. Eles não conseguiram, por exemplo, testar os efeitos duradouros, meses depois — para ver quanto tempo a queda na criminalidade permaneceria. E ainda é possível que a exposição constante a filmes violentos em algum momento leve a mais violência. Entretanto, o estudo conseguiu avaliar o impacto imediato desses filmes, que tem sido o principal tema dos experimentos. Talvez os filmes violentos influenciem as pessoas e as tornem excepcionalmente irritadas e agressivas. Entretanto, você sabe o que influencia de modo inegável as pessoas na direção da violência? Sair com outros homens potencialmente agressivos e beber.*

Agora isto faz sentido. Mas não fazia antes que Dahl e DellaVigna começassem a analisar pilhas de dados.

Um ponto mais importante que fica bastante claro quando analisamos dados: o mundo é complexo. As ações que praticamos hoje podem ter efeitos remotos, a maioria deles não intencional. As ideias se disseminam — às vezes lenta, outras exponencialmente, como vírus. As pessoas reagem de maneiras imprevisíveis a impulsos.

Essas conexões e relações, explosões e incrementos, não podem ser rastreadas com pesquisas pequenas ou métodos de dados tradicionais. O mundo, simplesmente, é complexo e rico demais para poucos dados.

NOSSOS DÚPLICES

Em junho de 2009, David “Big Papi” Ortiz parecia liquidado. Durante a década anterior, o Boston Red Sox havia se apaixonado por seu bateador de origem dominicana de sorriso amigável e lacuna entre os dentes.

Foi selecionado cinco vezes consecutivas para o jogo All-Star, eleito MVP e ajudou a encerrar o jejum de 86 anos sem vencer um campeonato do Boston Red Sox. Mas na temporada de 2008, aos 38 anos, seus números despencaram. Sua média de rebatimento caiu 68 pontos, seu percentual de bases conquistadas, 76, e seu percentual de *slugging*, 114. E no início da temporada de 2009, os números despencaram ainda mais.

Bill Simmons, jornalista esportivo e fã apaixonado do Boston Red Sox fan, descreveu o que aconteceu no início da temporada de 2009 da seguinte forma: “Está nítido que David Ortiz não é mais o mesmo… Batedores fortões são como estrelas pornôs, lutadores, pivôs da NBA e esposas troféu: quando despencam, despencam de uma vez.” Os grandes fãs do esporte confiam em seus olhos, e os de Simmons o diziam que Ortiz estava acabado. Na verdade, Simmons previu que seria relegado ao banco de reservas ou dispensado em breve.

Ortiz estava mesmo acabado? Se você fosse o diretor-geral do Red Sox, em 2009, o cortaria do time? De modo mais geral, como podemos prever o futuro desempenho de um jogador de beisebol? De modo ainda mais genérico, como podemos usar Big Data para prever o que as pessoas farão no futuro?

Uma teoria bastante produtiva na ciência de dados é a seguinte: analise o que os *sabermetricians* (aqueles que usam dados para estudar beisebol) fazem e espere que se espalhe para outras áreas da ciência de dados. O beisebol foi um dos primeiros campos com conjuntos de dados abrangentes sobre quase tudo, e um exército de pessoas inteligentes dispostas a devotar suas vidas a compreender os dados. Agora, praticamente todo os outros campos já chegaram ou se aproximam deste patamar. O beisebol foi o pioneiro, os outros o seguiram. A sabermétrica domina o mundo.

Uma forma bem simples de prever o futuro dos jogadores de beisebol é presumir que continuarão com o mesmo desempenho que têm. Se um jogador teve um baixo desempenho nos últimos 18 meses, você pode supor que

continuará assim pelos próximos 18 meses.

De acordo com esta metodologia, o Red Sox deveria cortar David Ortiz do time.

Entretanto, há informações mais relevantes. Na década de 1980, Bill James, considerado por muitos como o fundador da sabermétrica, enfatizou a importância da idade. Jogadores de beisebol, James descobriu, têm pico de desempenho ainda jovens — por volta dos 27 anos.

Os times tendem a ignorar o quanto o desempenho dos jogadores declina conforme envelhecem. Eles pagam caro demais por jogares que estão envelhecendo.

Por esta metodologia mais avançada, o Red Sox, com certeza, deveria ter cortado David Ortiz do time.

Mas esse ajuste de idade ignora algo a mais. Nem todos os jogadores seguem a mesma trajetória ao longo da vida. Alguns têm o ápice de seu desempenho aos 23 anos, outros, aos 31. Jogadores mais baixos envelhecem de forma diferente dos mais altos; os mais gordos, diferente dos mais magros. Os estatísticos do beisebol descobriram que há tipos de jogadores, cada um com determinado caminho de envelhecimento. Esta versão foi ainda pior para Ortiz: "batedores fortões", de fato, em média, têm o ápice de desempenho mais jovens, e declinam bruscamente depois dos 30 anos.

Se o Red Sox considerasse seu passado recente, sua idade e seu porte físico, deveria, sem dúvida, ter dispensado David Ortiz.

Então, em 2003, o estatístico Nate Silver introduziu um novo modelo, que batizou de PECOTA, para prever o desempenho de um jogador. Ele se mostrou o melhor — e, também, o mais arrojado de todos. Silver procurou por dúplices dos jogadores. Funciona da seguinte forma: crie uma base de dados de todos os jogadores da Liga Principal de Beisebol de todos os tempos, são mais de 18 mil homens. E inclua tudo que você sabe sobre eles: altura, idade, posição, médias de *home runs*, de rebatimento, *walks* e *strikeouts* para cada ano de suas carreiras. Agora, encontre os vinte jogadores que mais se pareçam com Ortiz até aquele momento de sua carreira — aqueles que jogaram como ele quando tinha 24, 25, 26, 27, 28, 29, 30, 31, 32 e 33 anos. Em outras palavras, encontre seus dúplices e depois veja como foram suas carreiras.

Uma busca por dúplice é outro exemplo de foco. Ela se concentra em um pequeno subconjunto de pessoas mais parecidas com determinada pessoa. E, assim como todo foco, quanto mais dados, melhor. Ocorre que os dúplices de Ortiz ofereceram uma previsão muito diferente para o futuro de Ortiz. Seus dúplices incluíam Jorge Posada e Jim Thome. Estes jogadores começaram suas carreiras mais devagar; tiveram explosões fantásticas aos vinte e tantos anos, com potencial para elite, e declinaram logo após os trinta.

Silver então previu como Ortiz se sairia com base nos resultados de seus dúplices. E descobriu que eles recuperaram o poder. Para as esposas troféus, Simmons pode estar certo: quando despencam, despencam de uma vez. Mas para os dúplices de Ortiz, quando despencaram, recuperaram a forma.

A busca por dúplices, a melhor metodologia usada para prever o desempenho de jogadores de beisebol, mostrou que o Red Sox deveria ter paciência com Ortiz. E o time, de fato, foi muito paciente com seu jogador "maduro". Em 2010, a média de Ortiz subiu para .270. Ele acertou 32 *home runs* e chegou ao time All-Star. Esse foi o primeiro de uma sequência de 4 jogos All-Star para Ortiz. Em 2013, rebatendo em sua posição tradicional, aos 37 anos, a média de rebatimento de Ortiz foi .688 na vitória do Boston Red Sox sobre o St. Louis, 4 games a 2 na Série Mundial. Ortiz foi eleito MVP da Série Mundial.* Assim que terminei a leitura da abordagem usada por Nate Silver para prever a trajetória de jogadores, imediatamente comecei a pensar se eu também não teria um dúplice.

As buscas por dúplices são promissoras em muitos campos, não apenas nos esportes. Será que conseguiria encontrar a pessoa que compartilha o maior número de interesses comigo? Se a encontrasse, poderíamos sair juntos. Ela deve conhecer algum restaurante que eu gostaria de frequentar. Talvez me apresentasse a coisas que não tenho a menor ideia de que tenho afinidade.

Uma busca por dúplices se concentra em indivíduos e até mesmo nas suas características. E, assim como com todo foco, quanto mais dados, mais preciso é o resultado. Suponha que eu busque meu dúplice em um conjunto de dados de mais ou menos dez pessoas. Posso encontrar alguém que compartilhe de meu interesse por livros. Imagine que eu busque meu dúplice em um conjunto de dados de mais ou menos mil pessoas. Posso encontrar alguém que tenha a mesma predileção por livros populares de física que eu. Mas suponha que eu procure meu dúplice em um conjunto de dados de centenas de milhões de pessoas. Neste caso, seria capaz de encontrar alguém realmente muito parecido comigo.

Um dia, saí à caça por um dúplice na rede social. Usando todo o conjunto de perfis do Twitter, procurei por uma pessoa no planeta que tivesse o maior número de interesses em comum comigo.

É possível descobrir muito sobre meus interesses a partir dos perfis que sigo em minha conta no Twitter. No total, sigo 250 pessoas, que mostram minhas paixões por esportes, política, comédia, ciência e cantores soturnos de música folclórica judaica.

Será que existe alguém no universo que siga os mesmos 250 perfis que eu, meu gêmeo do Twitter? Claro que não. Dúplices não são idênticos a nós, apenas parecidos. Também não há alguém que siga 200 ou 150 dos mesmos perfis.

Entretanto, descobri uma conta que segue impressionantes cem mesmos perfis que eu: *Country Music Radio Today*. Hein? Ocorre que, *Country Music Radio Today* era um robô (não existe mais) que seguia 750 mil perfis do Twitter na esperança de que o seguissem também.

Tenho uma ex-namorada que teria adoraria este resultado. Uma vez ela me disse que eu parecia mais um robô do que um ser humano.

Piadas à parte, minha descoberta inicial de que meu dúplice era um robô que seguia 750 mil contas aleatórias mostrou um ponto importante sobre as buscas por dúplices. Para que uma busca seja verdadeiramente precisa, você não quer encontrar alguém que simplesmente goste de tudo que gosta. Também quer achar quem deteste as mesmas coisas.

Meus interesses são evidentes não apenas pelas contas que sigo, mas também pelas que não sigo. Eu me interesso por esportes, política, comédia e ciência, e não por comida, moda ou teatro. Os perfis que sigo mostram que gosto de Bernie Sanders, mas não de Elizabeth Warren, curto Sarah Silverman, mas não Schumer, o *New Yorker*, mas não a *Atlantic*, meus amigos Noah Popp, Emily Sands e Josh Gottlieb, mas não meu amigo Sam Asher. (Desculpe, Sam. Mas seu Twitter me dá sono.)

De todas as 200 milhões de pessoas no Twitter, quem tem o perfil mais parecido com o meu? Meu dúplice, no fim das contas, é o escritor da *Vox* Dylan Matthews. Fiquei um pouco desapontado, para a finalidade de melhorar meu consumo de mídia, pois já sigo Matthews no Twitter e no Facebook, e leio compulsivamente seus posts na *Vox*. Assim, saber que ele é meu dúplice não mudou muito minha vida. Mas é legal saber qual é a pessoa mais parecida com você no mundo, especialmente se é alguém que admira. E quando terminar este livro e deixar de ser um ermitão, talvez Matthews e eu possamos sair para bater papo e discutir os trabalhos de James Surowiecki.

A busca pelo dúplice de Ortiz foi perfeita para os fãs de beisebol. A procura pelo meu dúplice foi divertida, pelo menos para mim. Mas o que mais essas buscas podem revelar? Primeiro, elas têm sido usadas pelas grandes empresas de internet para aprimorar drasticamente suas ofertas e a experiência do usuário. A Amazon usa algo parecido com a busca de dúplices para sugerir livros que possam lhe interessar. Eles analisam o que as pessoas parecidas com você escolheram e o utilizam para fazer recomendações.

O Pandora faz o mesmo ao escolher que músicas você pode gostar de ouvir. E é assim também que a Netflix descobre de que filmes você pode gostar. O impacto é tão grande que quando o engenheiro da Amazon, Greg Linden, originalmente introduziu a busca por dúplices para prever livros preferidos, o aprimoramento nas recomendações foi tão bom que o fundador da Amazon, Jeff Bezos, se ajoelhou e gritou para Linden: “Você é meu ídolo!”

Mas o que realmente é interessante sobre as buscas por dúplices, considerado todo seu potencial, não é como são normalmente usadas hoje. É com que frequência não o são. Existem aspectos importantes da vida que poderiam ser melhorados enormemente pelo tipo de personalização que essas buscas permitem.

Isaac Kohane, cientista de computação e médico pesquisador em Harvard, tenta implementar este princípio na medicina. Ele quer organizar e coletar toda nossa informação de saúde para que em vez de usar uma abordagem padronizada os médicos encontrem pacientes iguais a você. Assim, eles podem fazer diagnósticos e tratamentos personalizados e mais focados.

Kohane considera isso uma extensão natural do setor médico, e não é tão radical como pode parecer. “O que é um diagnóstico?” Kohane pergunta. “Um diagnóstico, na verdade, é uma declaração em que você compartilha características com populações previamente estudadas. Quando diagnostico alguém com um infarto, digo que tem uma patofisiologia que descobri a partir de outras pessoas, representando um ataque cardíaco.”

Um diagnóstico é, em essência, um tipo primitivo de busca de dúplice. O problema é que os conjuntos de dados usados pelos médicos para diagnosticar são pequenos. Atualmente um diagnóstico é baseado na experiência do médico com a população de pacientes que ele tratou, talvez suplementada por artigos acadêmicos de pequenas populações descobertas por outros pesquisadores. Como vimos, porém, para que uma busca por dúplices seja realmente frutífera é preciso incluir muitos casos a mais.

Esta é uma área que o Big Data realmente ajudaria. Então, por que demora tanto? Por que já não é amplamente utilizado? O problema está na coleta de dados. A maioria dos registros médicos ainda é mantida em papel, enterrada em pastas, e aqueles que são digitalizados geralmente têm formatos incompatíveis. Frequentemente temos melhores dados, observa Kohane, sobre beisebol do que sobre saúde. Mas medidas simples terão um enorme resultado. Kohane constantemente fala que é como colher “frutas em galhos baixos”. Ele acredita, por exemplo, que simplesmente criar um conjunto de dados completo dos gráficos de alturas e pesos das crianças e quaisquer doenças que possam ter seria revolucionário para os pediatras. O crescimento de cada criança seria então comparado aos de todas as outras. Um computador encontraria crianças com uma trajetória semelhante e automaticamente alertaria

sobre quaisquer padrões anormais. Ele poderia detectar prematuramente uma deficiência de altura, o que em certos cenários provavelmente indicaria uma de duas possíveis causas: hipotireoidismo ou tumor cerebral.

O diagnóstico precoce em ambos os casos seria uma enorme vantagem. “Esses são casos raros”, de acordo com Kohane, “1 a cada 10 mil casos. Crianças, de modo geral, são saudáveis. Acredito que poderíamos diagnosticá-las precocemente, pelo menos um ano mais cedo. Tenho absoluta certeza de que é possível.”

James Heywood é um empreendedor que tem uma abordagem diferente para lidar com problemas relacionados a conectar dados médicos. Ele criou um site, PatientsLikeMe.com [conteúdo em inglês], em que indivíduos relatam as próprias informações — suas condições, tratamentos e efeitos colaterais. Heywood já teve bastante sucesso em mapear os diferentes cursos que as doenças tomam e como se comparam com nosso entendimento comum da doença.

O objetivo de Heywood é recrutar um número suficiente de pessoas, abrangendo condições de saúde suficientes para que elas consigam encontrar seus dúplices de saúde. Heywood espera que você ache pessoas da mesma idade e gênero, com histórico e relato de sintomas semelhantes aos seus — e descubra o que funcionou no caso delas. Este seria um tipo muito diferente de medicina, sem dúvida.

HISTÓRIAS DOS DADOS

De muitas maneiras o ato de focar a pesquisa é mais valioso para mim do que descobertas específicas de um estudo em particular, porque oferece uma nova forma de ver e falar sobre a vida.

Quando as pessoas descobrem que sou cientista de dados e escritor, às vezes, me contam algum fato ou comentam sobre uma pesquisa. Normalmente acho este tipo de dado maçante — estático e sem vida. Não há história sendo contada.

Do mesmo modo, amigos já tentaram fazer com que eu leia romances e biografias de que gostaram. Mas isto também não me atrai. Sempre acabo me perguntando: “O que aconteceria em outras situações? Qual seria o princípio mais geral?” Estas histórias parecem pequenas e sem representatividade.

O que mostro neste livro é algo que, para mim, é inigualável. É baseado em dados e números; ilustrativo e muito poderoso. E ainda assim os dados são tão ricos que é possível visualizar as pessoas por trás deles. Quando focamos cada minuto do consumo de água de Edmonton, *vejo* as pessoas se levantando da poltrona no final de cada tempo. Quando nos concentramos nas pessoas mudando da Filadélfia para Miami e começando a fraudar os impostos, as *vejo* conversando em suas casas e descobrindo o truque. Quando observamos os fãs de beisebol de cada idade, *vejo* minha própria infância e a de meus irmãos e a de milhões de homens adultos chorando pela derrota do time quando tinham 8 anos.

Correndo novamente o risco de parecer arrogante, acho que os economistas e os cientistas de dados mostrados neste livro estão criando não apenas uma nova ferramenta, mas uma nova espécie. O que tentei apresentar até agora neste capítulo, e em grande parte deste livro, é que os dados são tão grandes e tão ricos, nos permitindo focar de modo tão preciso que, sem nos limitarmos a um ser humano específico e não representativo, podemos ainda contar histórias complexas e envolventes.

* Informação completa: Quando checava as informações deste livro, Noah negou que sua aversão pelo passatempo preferido dos norte-americanos seja uma parte central de sua identidade. Ele admite odiar o beisebol, mas acredita que sua gentileza, amor pelas crianças e inteligência são os elementos essenciais de sua personalidade — e que sua opinião sobre beisebol não está nem mesmo entre seus dez traços principais. Entretanto, concluí que às vezes é difícil vermos a própria identidade objetivamente e, como um observador externo, sou capaz de notar que odiar beisebol é de fato fundamental para sua identidade, seja ele capaz de reconhecer ou não. Então, mantive a informação.

* Esta história mostra como as coisas que parecem más podem ser, de fato, boas se prevenirem o pior. Ed McCaffrey, formado em Stanford e ex-jogador de futebol americano, usa esse argumento para justificar que seus quatro filhos joguem futebol americano: “Eles têm energia. E, assim, se não estão jogando futebol, estão andando de skate, subindo em árvores, brincando de pega-pega no quintal, jogando paintball. Quero dizer, eles não vão ficar sentados fazendo nada. A forma de encarar isso é, bem, pelo menos no esporte existem regras… Meus filhos já foram parar no pronto-socorro por cair de deques, tombos de bicicletas, skates, cair de árvores. Tudo que você possa imaginar… Sim, é um esporte de impactos violentos. Mas, também, meus garotos têm personalidade, e, pelo menos eles não estão saltando de montanhas com trajes voadores ou fazendo outras maluquices. Então, é uma agressão organizada… eu acho.” Eu nunca tinha ouvido antes este tipo de argumento de McCaffrey, de uma entrevista para *The Herd with Colin Cowherd*. Depois de ler o artigo de Dahl/DellaVigna, levo esta alegação muito a sério. Uma vantagem dos conjuntos de dados gigantescos do mundo real, contra os dados de laboratório, é que conseguem captar estes tipos de efeito.

* Você já deve ter percebido neste ponto do livro que tendo a ser cínico sobre boas histórias. Mas queria ao menos uma história feliz aqui, então deixei meu cinismo para a nota de rodapé. Desconfio que o PECOTA acaba de descobrir que Ortiz usava esteroides, parou e depois voltou a usar. Do ponto de vista da previsão, na verdade, seria fantástico se o PECOTA fosse capaz de detectar este tipo de coisa — mas a história perde o toque comovente.

6

O MUNDO TODO É UM LABORATÓRIO

O dia 27 de fevereiro de 2000 começou como um dia qualquer no campus de Mountain View, da Google. O sol estava brilhando, ciclistas, pedalando, massagistas, massageando, os funcionários, se hidratando com água aromatizada de pepino. E então, neste dia absolutamente normal, alguns engenheiros da Google tiveram uma ideia que desvendou o segredo que hoje movimenta grande parte da internet. Os engenheiros descobriram a melhor maneira de fazer você clicar, voltar ou passar mais tempo nos sites da Google.

Antes de descrever o que fizeram, precisamos falar sobre correlação versus causalidade, uma grande questão na análise de dados — que ainda não discutimos adequadamente.

A mídia nos bombardeia com estudos baseados em correlação aparentemente todos os dias. Por exemplo, dizem que quem bebe uma quantidade moderada de álcool tende a ser mais saudável. Isto é correlação.

Isto significa que beber uma quantia moderada de álcool melhorará sua saúde — uma relação de causalidade? Talvez, não. Pode ser que a boa saúde faça com que as pessoas bebam moderadamente. Cientistas sociais chamam isto de causalidade reversa. Ou pode haver um fator independente que provoque as duas coisas: boa saúde e beber com moderação. Talvez passar bastante tempo com amigos leve tanto a um consumo moderado de álcool quanto à boa saúde. Cientistas sociais chamam isso de viés de variável omitida.

Como, então, estabelecemos causalidade de modo mais preciso? O método mais confiável é um experimento aleatório controlado. Veja como funciona. Divide-se as pessoas aleatoriamente em dois grupos. A um deles, o grupo de tratamento, pede-se que tome ou faça determinada coisa. O outro, o grupo de controle, não faz nada. Então, observa-se como cada um deles responde. A diferença nos resultados entre os dois grupos é o efeito causal.

Por exemplo, para testar se o consumo moderado de álcool é a causa da boa saúde, você pode escolher aleatoriamente algumas pessoas para beber uma taça de vinho por dia e outras para não beber álcool, por um ano, e depois comparar os relatos de saúde dos dois grupos. Como as pessoas foram escolhidas arbitrariamente para cada grupo, não há razão para esperar que um grupo tenha melhor saúde inicial ou socializado mais. Você pode confiar que os efeitos do vinho são causais. Experimentos aleatórios controlados são a evidência mais confiável em qualquer campo. Se um medicamento passar em um experimento assim, pode ser distribuído para o público. Se não passar no teste, não chega às prateleiras das farmácias.

Experimentos randômicos vêm sendo cada vez mais usados também por cientistas sociais. Esther Duflo, economista francesa do MIT, liderou a campanha pela maior utilização dos experimentos em economia desenvolvimental, um campo que tenta descobrir as melhores maneiras de ajudar as pessoas mais pobres no mundo. Vejamos o estudo de Duflo e seus pares sobre como melhorar a educação na área rural da Índia, em que mais da metade dos alunos de ensino médio não sabe ler uma frase simples. Um dos potenciais motivos para os alunos terem tanta dificuldade é a constante ausência dos professores. Em um dia qualquer em algumas escolas da área rural da Índia, mais de 40% dos professores não comparece.

O teste de Duflo? Ela e seus colegas dividiram aleatoriamente as escolas em dois grupos. Um (o grupo de tratamento), além do pagamento normal, recebia uma pequena quantia — 50 rúpias, ou aproximadamente US$1,15 — para cada dia que comparecessem ao trabalho. Ao outro não foi feito qualquer pagamento adicional. Os resultados foram visíveis. Quando os professores recebiam, a ausência caía pela metade. O teste de desempenho dos alunos também melhorou substancialmente, com maiores efeitos sobre as meninas. Ao final do experimento, as garotas nas escolas em que os professores foram pagos para comparecer às aulas tiveram uma probabilidade 7 pontos percentuais maior de saber escrever.

De acordo com um artigo da revista *New Yorker*, quando Bill Gates soube do trabalho de Duflo, ficou tão impressionado que disse a ela: “Nós *precisamos* patrocinar você.”

OS FUNDAMENTOS DO TESTE A/B

Os experimentos aleatórios são o método mais confiável para provar causalidade, e sua utilização se disseminou pelas ciências sociais. Isto nos leva de volta aos escritórios da Google, em 27 de fevereiro de 2000. O que a Google fez que revolucionou a internet?

Nesse dia, alguns engenheiros decidiram realizar um experimento no site do Google. Dividiram aleatoriamente os usuários em dois grupos. O grupo de tratamento recebia vinte links nas páginas de resultados de busca. O grupo de controle, os dez links como de costume. Os engenheiros então compararam a satisfação dos dois grupos com base na frequência com que retornavam ao Google.

Isto é revolução? Não me parece tão vanguardista. Já mencionei que experimentos aleatórios têm sido usados por companhias farmacêuticas e cientistas sociais. Como a simples imitação desses métodos pode ser tão relevante?

O ponto principal — e isto foi rapidamente percebido pelos engenheiros da Google — é que os experimentos no mundo digital têm uma gigantesca vantagem em relação ao mundo offline. Por mais convincentes que possam ser os experimentos aleatórios offline, também requerem muitos recursos. Para o estudo de Duflo, as escolas tiveram que ser contatadas, financiamentos, obtidos, alguns professores tiveram que ser pagos e todos os alunos, testados. Experimentos offline custam milhares ou centenas de milhares de dólares e levam meses para ser conduzidos.

No mundo digital, experimentos aleatórios podem ser baratos e rápidos. Você não precisa recrutar e pagar participantes. Basta escrever uma linha de código para atribuir arbitrariamente pessoas a um grupo. Não é preciso que os usuários respondam questionários de pesquisa. Em vez disso, você pode medir os movimentos e cliques do mouse. Não é preciso codificar manualmente e analisar respostas. Basta criar um programa para fazer isto automaticamente. Não precisa contatar ninguém. Nem mesmo precisa contar aos usuários que estão participando de um experimento.

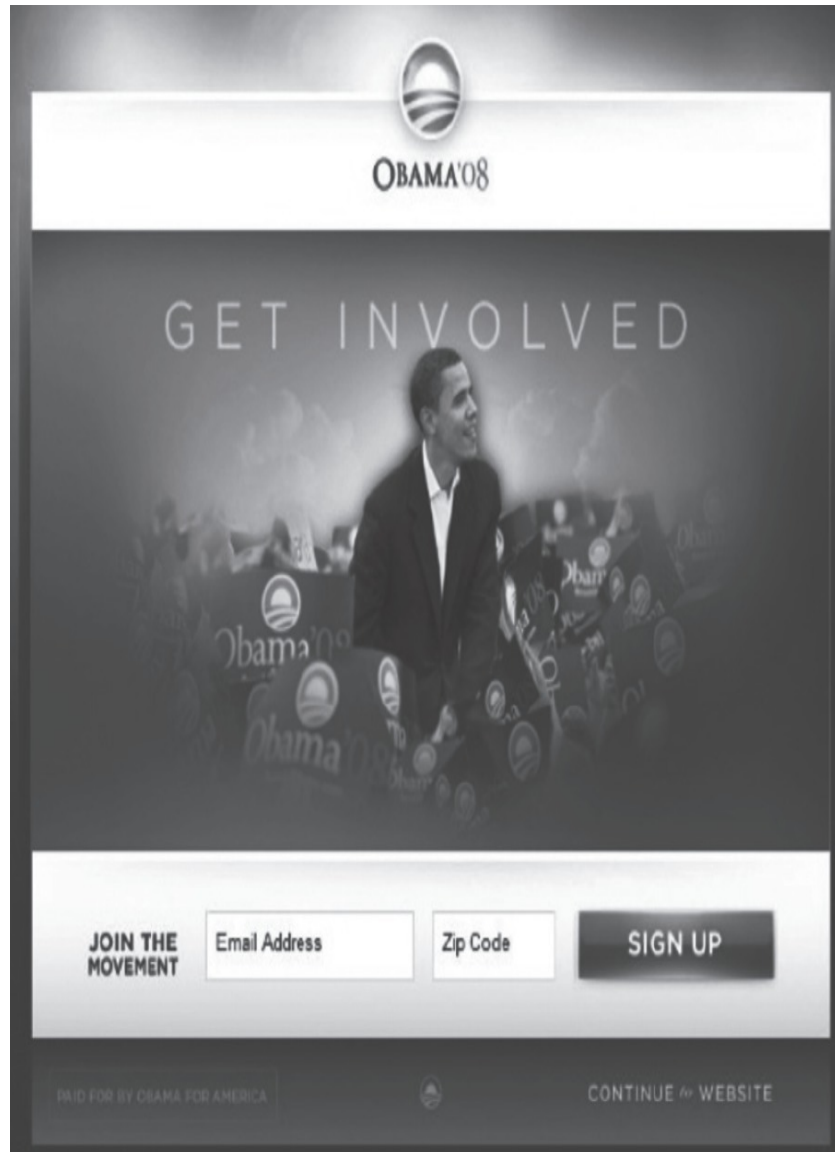
Este é o quarto poder do Big Data: fazer experimentos aleatórios, capazes de descobrir efeitos causais autênticos, muito, muito mais fáceis de ser conduzidos — a qualquer hora, mais ou menos em qualquer lugar, desde que você esteja online. Na era do Big Data, o mundo todo é um laboratório.

Esta nova perspectiva rapidamente se alastrou pela Google e depois pelo restante do Vale do Silício, em que experimentos randômicos controlados foram rebatizados de “teste A/B”. Em 2011, os engenheiros da Google executaram 7 mil deles. E este número não para de crescer.

Se a Google quer fazer com que mais pessoas cliquem nos anúncios em seus sites, pode experimentar dois tons de azul nos anúncios — um para o Grupo A, outro para o B. A Google então compara as taxas de cliques. Obviamente, a facilidade deste teste leva a exageros. Alguns funcionários acharam que, em razão da facilidade, a Google estava extrapolando. Em 2009, um designer frustrado se demitiu depois que a Google submeteu 41 tons levemente diferentes de azul ao teste A/B. Mas a tentativa de defender a arte contra a obsessiva pesquisa de mercado desse designer teve pouco efeito para impedir a disseminação desta metodologia.

O Facebook hoje conduz mil testes A/B por dia, o que significa que um pequeno número de engenheiros no Facebook inicia mais experimentos aleatórios controlados em um dia do que a indústria farmacêutica inteira em um ano.

O teste A/B se disseminou além das grandes empresas de tecnologia. Um ex-funcionário da Google, Dan Siroker, levou a metodologia para a primeira campanha presidencial de Barack Obama, em que utilizou o teste A/B nos designs dos sites, no discurso dos e-mails e nos formulários de doações. Depois Siroker criou uma nova empresa, a Optimizely, que permite que organizações conduzam testes A/B rápidos. Em 2012, a Optimizely foi utilizada por Obama, e também por seu oponente, Mitt Romney, para maximizar o número de assinaturas, voluntários e doações. A empresa também é utilizada por uma ampla gama de companhias, como Netflix, TaskRabbit e a revista *New York*.

Para avaliar como este teste é valioso, vejamos como Obama o utilizou para engajar mais pessoas em sua campanha. O site da campanha de Obama inicialmente incluiu uma foto do candidato e um botão logo abaixo, com o convite “Sign up” [Assine].

Esta era a melhor maneira de saudar as pessoas que acessavam o site? Com a ajuda de Siroker, a equipe de Obama testava se uma foto e um botão diferente faziam com que mais pessoas assinassem o site. Será que mais pessoas clicariam se ele exibisse uma foto de Obama com a expressão mais séria? Clicariam mais se no botão estivesse escrito “Join Now” [Junte-se a Nós Agora]? A equipe de Obama mostrou aos usuários diferentes combinações de fotos e botões e avaliou quantos deles clicaram. Veja se consegue adivinhar a foto e o botão vencedores.

Fotos testadas

Botões Testados [Saiba Mais; Assine; Junte-se a Nós]

A vencedora foi a foto da família Obama e o botão “Learn More” [Saiba Mais]. E a vitória foi esmagadora. Usando esta combinação, a equipe de campanha de Obama estimou que a associação fez com que 40% mais pessoas assinassem o site, rendendo aproximadamente US$60 milhões em doações adicionais.

Combinação Vencedora

Há outro grande benefício além do fato de todos estes testes tão confiáveis poderem ser feitos de modo barato e fácil: nos liberta ainda mais de confiar apenas na intuição, o que, como mencionado no Capítulo 1, tem suas limitações. Um motivo essencial para a importância do teste A/B é que as pessoas são imprevisíveis. Nossa intuição frequentemente é incapaz de prever como vão responder.

Sua intuição estava correta sobre o site ideal para Obama?

Veja mais alguns testes intuitivos. Analise os testes A/B de manchetes do *Boston Globe* para descobrir quais fazem as pessoas clicar mais em um artigo. Tente adivinhar os vencedores entre estas opções:

UMA DESTAS MANCHETES SE SAIU MUITO MELHOR QUE A OUTRA EM OBTER CLIQUES.

	MANCHETE A	MANCHETE B
1.	O drone SnotBot é capaz de salvar as baleias?	Este drone é capaz de salvar as baleias?
2.	Obviamente "bolas murchas" é um termo de busca popular em Massachusetts	O termo de busca no Google mais popular em Massachusetts é bastante embaraçoso
3.	Circo da mídia domina o julgamento do estupro em St. Paul	Não houve denúncia no escândalo sexual da escola fundamental
4.	Mulher ganha uma bolada com raro cartão de beisebol	Mulher ganha U$179.000 com raro cartão de beisebol
5.	MBTA projeta que o deficit de operação anual duplicará até 2020	Prepare-se: o deficit da MBTA está prestes a dobrar
6.	Como Massachusetts ajudou você a conquistar o direito ao acesso a métodos de controle de natalidade	Como a Universidade de Boston ajudou a acabar com os "crimes contra a castidade"
7.	Quando inauguraram o primeiro metrô em Boston	Cartoons de quando o primeiro metrô de Boston foi inaugurado
8.	Vítima e familiares no caso de estupro na escola fundamental culpam cultura tóxica	Vítima e familiares no caso do estupro na escola fundamental divulgam declaração
9.	Homem usando boné escrito "Free Brady" é o único capaz de descobrir disfarce de Miley Cyrus	Fã de Pats vira centro das atenções depois de reconhecer Miley Cyrus disfarçada

Fez seus palpites? As respostas estão em negrito na página seguinte.

	MANCHETE A	MANCHETE B	VENCEDOR?
1.	**O drone SnotBot é capaz de salvar as baleias?**	Este drone é capaz de salvas as baleias?	53% mais cliques para A
2.	Obviamente “bolas murchas” é um termo de busca popular em Massachusetts	**O termo de busca no Google mais popular em Massachusetts é bastante embaraçoso**	986% mais cliques para B
3.	Circo da mídia domina o julgamento do estupro em St. Paul	**Não houve denúncia no escândalo sexual da escola fundamental**	108% mais cliques para B
4.	**Mulher ganha uma bolada com raro cartão de beisebol**	Mulher ganha U$179.000 com raro cartão de beisebol	38% mais cliques para A
5.	MBTA projeta que o deficit de operação anual duplicará até 2020	**Prepare-se: o deficit da MBTA está prestes a dobrar**	62% mais cliques para B
6.	Como Massachusetts ajudou você a conquistar o direito ao acesso a métodos de controle de natalidade	**Como a Universidade de Boston ajudou a acabar com os “crimes contra a castidade”**	188% mais cliques para B
7.	**Quando inauguraram o primeiro metrô em Boston**	Cartoons de quando o primeiro metrô de Boston foi inaugurado	33% mais cliques para A
8.	Vítima e familiares no caso de estupro na escola fundamental culpam cultura tóxica	**Vítima e familiares no caso do estupro na escola fundamental divulgam declaração**	76% mais cliques para B
9.	Homem usando boné escrito “Free Brady” é o único capaz de descobrir disfarce de Miley Cyrus	**Fã de Pats vira centro das atenções depois de reconhecer Miley Cyrus disfarçada**	67% mais cliques para B

Prevejo que você acertou mais da metade, talvez avaliando em qual deles clicaria. Mas provavelmente não conseguiu adivinhar todas corretamente.

Por quê? O que você deixou passar? Quais percepções sobre o comportamento humano lhe faltam? Que lições você pode aprender a partir desses erros?

Normalmente nós fazemos perguntas como essas depois de previsões erradas.

Mas veja como é difícil chegar a conclusões genéricas a partir das manchetes do *Globe*. No teste da primeira manchete, mudar uma única palavra, “este” por “SnotBot”, apresentou um bom resultado. Isso pode sugerir que oferecer mais detalhes é melhor. Mas na segunda manchete, “bolas murchas”, o termo detalhado, é o menos clicado. Na quarta manchete, “ganhou uma bolada” superou o número US$179.000. Isso pode indicar que gírias dão melhores resultados. Mas a gíria “circo da mídia” perde na terceira manchete.

A lição do teste A/B, em grande extensão, é desconfiar de regras gerais. Clark Benson é CEO do ranker.com, um site de notícias e entretenimento que confia solidamente no teste A/B para escolher as manchetes e o design do site. “No fim das contas, não se pode presumir nada”, diz Benson. “Teste literalmente tudo.”

Os testes preenchem as lacunas em nossa compreensão da natureza humana. Essas lacunas sempre irão existir. Se soubéssemos, com base em nossa experiência de vida, qual seria a resposta, os testes não teriam valor. Mas não sabemos.

Outra razão do teste A/B é tão importante que aparentemente pequenas mudanças podem ter grandes efeitos. Como Benson coloca: “Constantemente me impressiono como fatores tão pequenos se mostram tão valiosos nos testes.”

Em dezembro de 2012, a Google mudou seus anúncios. Ela acrescentou uma seta apontando para a direita dentro de um quadrado.

Hotels
www.example.com
Special rates until the end of the month. No booking fees, book your room now!

Dublin Hotels
www.example.com
Browse hundreds of hotels in Dublin, sort by price, location and user reviews.

Hotels in Ireland
www.example.com
Compare prices of 1000s of hotels all over ireland!

AdChoices

Veja que seta bizarra. Ela aponta para absolutamente nada. Na verdade, a primeira vez que as setas apareceram, muitos usuários do Google criticaram. Por que eles estavam acrescentando uma seta sem sentido em seus anúncios, se perguntaram?

Bem, o Google é muito cuidadoso com seus segredos comerciais, então não diz exatamente qual o valor dessas setas, mas declararam que elas venceram no teste A/B. A Google adicionou as setas porque elas aumentaram o número de cliques. E essa pequena e aparentemente insignificante mudança rendeu à Google e suas parceiras montanhas de dinheiro.

Então, como podemos descobrir esses pequenos ajustes que produzem imensos benefícios? É preciso testar muitas coisas, mesmo que muitas pareçam triviais. Na verdade, os usuários do Google inúmeras vezes percebem minúsculas mudanças nos anúncios que depois voltam a ser como antes. Eles involuntariamente se tornaram membros de grupos de tratamento de testes A/B — mas com o único inconveniente de ver essas pequenas variações.

Experimento de Centralização (Não Funcionou)

Best Selling iPad 2 Case

The ZAGGmate™ - Tough Aluminum Case with build in Bluetooth Keyboard

www.zagg.com

Experimento da Estrela Verde (Não Funcionou)

Foster's Hollywood Restaurant **Reviews**, Madrid, Spain ...

www.tripadvisor.co.uk › ... › Madrid › Madrid Restaurants ▾ TripAdvisor ▾

★★★★★ Rating: 3 - 118 reviews

Foster's Hollywood, Madrid: See 118 unbiased **reviews** of **Foster's Hollywood**, rated 3 of 5 on TripAdvisor and ranked #3647 of 6489 restaurants in Madrid

Experimento de Nova Fonte (Não Funcionou)

Live Stock Market News

Free Charts, News and Tips from UTVi Experts Today!

UTVi.com/Stocks

As variações anteriores nunca foram exibidas para o público geral. Elas perderem no teste. Mas fizeram parte de um processo de escolha de vencedores. O caminho até a seta clicável é repleto de estrelas horrorosas, posicionamentos errados e fontes chamativas.

Pode ser divertido adivinhar o que faz as pessoas clicarem. E se você for um Democrata, talvez goste de saber que utilizar os testes rendeu a Obama muito dinheiro. Mas existe um lado sombrio do teste A/B.

Em seu excelente livro *Irresistible* [“Irresistível”, em tradução livre], Adam Alter escreve sobre o aumento dos vícios comportamentais na sociedade contemporânea. Muitas pessoas estão descobrindo aspectos da internet difíceis de desligar.

Meu conjunto de dados favorito, as buscas no Google, pode nos dar algumas pistas do que as pessoas acham mais viciantes. De acordo com o Google, a maioria dos vícios continuam sendo aqueles com que as pessoas lutam há muitas décadas — drogas, sexo e álcool, por exemplo. Mas a internet está começando a marcar presença na lista — com “pornografia” e “Facebook” agora entre os dez vícios mais reportados.

PRINCIPAIS VÍCIOS RELATADOS AO GOOGLE EM 2016

Drogas	Álcool	Jogos de azar
Sexo	Açúcar	Facebook
Pornografia	Amor	

O teste A/B pode estar desempenhando um papel significativo em tornar a internet tão viciante.

Tristan Harris, "especialista em ética em design", foi citado em *Irresistible* explicando por que as pessoas têm tanta dificuldade em resistir a determinados sites da internet: "Existem milhares de pessoas do outro lado da tela cujo trabalho é vencer seus mecanismos de autocontrole."

E essas pessoas usam o teste A/B.

Através dos testes, o Facebook pode descobrir que mudar um botão para determinada cor faz com que as pessoas voltem ao site com mais frequência. Então, mudam a cor do botão. Depois, podem descobrir que uma determinada fonte traz o mesmo resultado. E mudam a fonte também. Então, descobrem que enviar e-mails para os usuários em determinados horários também torna os usuários mais assíduos. E passam a enviar o e-mail.

Muito em breve, o Facebook está otimizado em maximizar o tempo que as pessoas passam em seu site. Em outras palavras, descubra um número suficiente de vencedores em testes A/B e você terá um site viciante. Esse é um tipo de feedback que as companhias de cigarros nunca tiveram.

O teste A/B está se tornando cada vez mais uma ferramenta da indústria de jogos. Como alega Alter, o *World of Warcraft* executa testes A/B em diversas versões do jogo. Uma missão pode pedir que o jogador mate alguém. Outra, que você salve alguma coisa. Os designers do jogo podem oferecer diferentes amostras de missões de jogadores e depois avaliar qual delas mantém mais pessoas jogando. Eles podem descobrir, por exemplo, que a missão que pedia que o jogador salvasse uma pessoa fez com que as pessoas retornassem com 30% mais frequência. Se testarem muitas missões, encontram mais e mais vencedores. Esses 30% de vantagem vai se somando, até que eles tenham um jogo capaz de manter muitos homens adultos enfurnados no porão da casa dos pais.

Se você está um pouco perturbado com essa informação, somos dois. E falaremos um pouco mais sobre as implicações éticas desse e de outros aspectos do Big Data mais adiante neste livro. Mas para o bem ou para o mal, a experimentação é agora uma ferramenta crucial à disposição dos cientistas de dados. E existe uma nova forma de experimentação nesta caixa de ferramentas. Ela tem sido usada para fazer uma variedade de questionamentos, incluindo se os anúncios de TV realmente funcionam.

EXPERIMENTOS CRUÉIS — MAS ESCLARECEDORES — DA NATUREZA

No dia 22 de janeiro de 2012, o New England Patriots jogava contra o Baltimore Ravens em uma partida do Campeonato AFC.

Falta um minuto para o fim do jogo. O Ravens está perdendo, mas tem a posse de bola. Os próximos sessenta segundos irão determinar qual time disputará o Super Bowl; ajudarão a selar o legado desses jogadores. E o último minuto do jogo decidirá algo que, para um economista, é muito mais profundo: os sessenta segundos finais ajudarão a nos dizer, de uma vez por todas, se as propagandas realmente funcionam.

A noção de que os anúncios aumentam as vendas é obviamente essencial para nossa economia. Mas isso é insanamente difícil de provar. Na verdade, esse é um exemplo clássico do quanto é difícil distinguir entre correlação e causa.

Não há dúvida de que os produtos que fazem mais propagandas são também os que vendem mais. A Twentieth Century Fox gastou US$150 milhões no marketing do filme *Avatar*, que se tornou o filme de maior bilheteria de todos os tempos. Mas quanto dos US$2,7 bilhões das vendas dos ingressos de *Avatar* foi provocada pelo marketing pesado? Parte do motivo da 20th Century Fox ter gasto tanto dinheiro na promoção supostamente era saberem que tinham um produto de grande apelo.

As empresas acreditam que sabem o quanto seus anúncios são eficientes. Os economistas são céticos. O professor de economia da Universidade de Chicago, Steven Levitt, quando colaborou com uma empresa de

eletrônicos, não se impressionou quando tentaram lhe convencer de que sabiam o grau de eficácia de seus anúncios. Como, Levitt pensou, eles podiam ser tão confiantes?

A empresa explicou que, todo ano, nos dias que antecedem o Dia dos Pais, eles aumentam seus gastos em propaganda de TVs. Como previsto, todo ano, antes do Dia dos Pais, eles tinham um pico de vendas. Ah, talvez isso seja apenas porque muitos filhos comprem eletrônicos para os pais, especialmente como presente do Dia dos Pais, independentemente da propaganda.

"Eles inverteram completamente a causalidade", disse Levitt em uma palestra. "Ao menos é o que parece. Não sabemos. Na verdade, esse é o problema", acrescenta Levitt.

Por mais importante que seja resolver este problema, as empresas relutam em conduzir experimentos rigorosos. Levitt tentou convencer as empresas de eletrônicos a realizar um experimento aleatório controlado para saber com precisão o quanto seus anúncios são eficientes. Como ainda não é possível fazer o teste A/B pela televisão, seria preciso analisar o que acontece quando não se faz propaganda em algumas áreas.

A empresa respondeu: "Você está maluco? Não podemos deixar de fazer propaganda em todos os meios. O CEO nos mataria." Esse foi o fim da colaboração de Levitt na empresa.

O que nos leva de volta ao jogo do Patriots versus Ravens. Como os resultados de um jogo de futebol americano podem nos ajudar a determinar os efeitos causais da propaganda? Bem, ele não é capaz de nos mostrar os efeitos de uma campanha de propaganda específica de uma determinada empresa. Mas pode nos oferecer evidências nos efeitos médios das propagandas de muitas grandes empresas.

Ocorre que, há um experimento de propaganda oculto em jogos como este. Veja como funciona. No momento que essas partidas são disputadas, as empresas já compraram, e produziram os anúncios do Super Bowl. Quando elas decidem que anúncios criar, não sabem quais times irão disputar a partida.

Mas os resultados das finais dos campeonatos da Confederação Americana de Futebol e da Confederação Nacional de Futebol, que determinam quais times disputarão o Super Bowl, têm um gigantesco impacto em quem de fato assistirá a partida. Os dois times que se qualificarem trarão uma gigantesca horda de telespectadores. Se o Patriots, que joga perto de Boston, vencer, mais pessoas da região de Boston e não de Baltimore vão assistir ao jogo. E vice-versa.

Para as empresas, isso é o equivalente a jogar cara ou coroa para determinar se dezenas de milhares de pessoas em Baltimore ou em Boston serão expostas a seus anúncios, e o resultado dessa aposta só será conhecido depois que os tempos de anúncios e os comerciais já tiverem sido produzidos.

Agora, voltando ao campo, onde Jim Nantz da CBS está anunciando os resultados deste experimento.

> Lá vem Billy Cundiff, para empatar o jogo e, ao que parece, teremos uma prorrogação. Nos dois últimos anos, dezesseis a dezesseis field goals.* Trinta e duas jardas para o empate. E o chute. Lá vai! Lá vai! E não deu… E o Patriots finalizam a jogada e compram sua passagem para Indianápolis. Eles estão no 46° Super Bowl.

Duas semanas depois, o Super Bowl XLVI atingiria 60,3 pontos de audiência em Boston e 50,2 pontos em Baltimore. Sessenta mil pessoas a mais em Boston assistiriam às propagandas de 2012.

No ano seguinte, os mesmos dois times se encontrariam no Campeonato da AFC. Desta vez, Baltimore venceria. A audiência adicional para os anúncios do Super Bowl de 2013 seria de Baltimore.

	AUDIÊNCIA DO SUPER BOWL DE 2012 (REGIÃO DE BOSTON JOGANDO)	AUDIÊNCIA DO SUPER BOWL DE 2013 (REGIÃO DE BALTIMORE JOGANDO)
Boston	56,7	48,0
Baltimore	47,9	59,6

Hal Varian, economista chefe da Google; Michael D. Smith, economista da Carnegie Mellon; e eu usamos esses dois jogos e todos os outros Super Bowls de 2004 a 2013 para testar se — e, em caso positivo, quanto — os anúncios no Super Bowl funcionam. Especificamente, analisamos se quando uma empresa anuncia um filme no Super Bowl, percebem um salto na venda de ingressos nas cidades que tiveram maior índice de audiência do jogo.

E, de fato, tiveram. As pessoas nas cidades dos times que se qualificaram para o Super Bowl assistiram aos filmes anunciados durante o Super Bowl em uma taxa significativamente mais alta do que nas cidades dos times que perderam a partida de qualificação. Mais pessoas naquelas cidades viram o comercial. Mais pessoas naquelas cidades decidiram assistir ao filme.

Uma explicação alternativa pode ser que ver seu filme participando de um Super Bowl deixe as pessoas mais dispostas a ir ao cinema. Entretanto, testamos um grupo de filmes com orçamentos semelhantes e que foram lançados em momentos parecidos, mas que não foram anunciados no Super Bowl. Não houve um aumento de bilheteria nas cidades dos times que jogaram o Super Bowl.

Tudo bem, como você deve ter adivinhado, propaganda funciona. Isso não é tão surpreendente.

Mas a questão não é apenas o fato de darem resultado. Os anúncios foram incrivelmente eficientes. Na verdade, na primeira vez que vimos os resultados, checamos duas, três, quatro vezes para garantir que estavam corretos — pois o retorno era muito grande. A média que um filme de nossa amostra pagou por um anúncio no Super Bowl foi cerca de US$3 milhões. O aumento de ingressos rendeu US$8,3 milhões, uma taxa de retorno sobre investimento de 2,8 para 1.

Este resultado foi confirmado por outros dois economistas, Wesley R. Hartmann e Daniel Klapper, que anteriormente e de forma independente tiveram ideia semelhante. Esses economistas estudaram anúncios de cerveja e de refrigerantes exibidos durante o Super Bowl, também utilizando o aumento da exposição aos anúncios nas cidades dos times que se qualificaram. Eles descobriram um retorno sobre investimento de 2,5 para 1. Por mais caros que esses anúncios sejam, nossos resultados e os deles sugerem que eles são tão eficientes em aumentar a demanda que as empresas acabam pagando muito barato, em razão do lucro obtido.

E o que tudo isso significa para nossos amigos da empresa de eletrônicos com que Levitt trabalhou? É possível que os anúncios no Super Bowl sejam mais lucrativos do que outras formas de propaganda. Mas na pior das hipóteses nosso estudo indica que toda aquela propaganda no Dia dos Pais provavelmente é sim uma boa ideia.

Uma vantagem do experimento do Super Bowl é que não foi necessário atribuir intencionalmente as pessoas para grupos de tratamento ou de controle. Ele ocorreu baseado nos caprichos da sorte em um jogo de futebol americano. Em outras palavras, aconteceu naturalmente. Por que isso é uma vantagem? Porque experimentos aleatórios controlados não naturais, apesar de superpoderosos e mais fáceis de conduzir na era digital, ainda nem sempre são possíveis.

Às vezes, não conseguimos fazer nossa mágica a tempo. Por vezes, assim como no caso da empresa de eletrônicos que não quis conduzir um experimento em sua campanha, estamos muito certos da resposta para a testar.

Em algumas ocasiões os experimentos são impossíveis. Suponha que você esteja interessado em quanto um país reage à perda de um líder. Ele vai a guerra? Sua economia para de funcionar? Não há grandes mudanças? Obviamente, não podemos simplesmente matar um número significativo de presidentes e primeiros-ministros para ver o que acontece. Isso não seria apenas impossível, mas também imoral. As universidades criaram, ao longo de muitas décadas, conselhos de revisão institucional (IRB, da sigla em inglês) que determinam se um experimento proposto é ético.

Então se quisermos saber os efeitos causais em um determinado cenário e se um experimento é antiético ou inviável, o que podemos fazer? Podemos utilizar o que os economistas — definindo a natureza de modo tão abrangente que inclui jogos de futebol — chamam de experimentos naturais.

Para melhor ou para pior (tudo bem, claramente para pior), existe um enorme componente de aleatoriedade na vida. Ninguém sabe ao certo o que ou quem controla o universo. Mas uma coisa está clara: quem quer que seja que comande o show — as leis da mecânica quântica, Deus, um garoto cheio de espinhas de cuecas simulando o universo em seu computador — elas, Ele ou ele *não está* sujeito à aprovação do IRB.

A natureza experimenta conosco o tempo todo. Duas pessoas levam um tiro. Uma bala para a milímetros de um órgão vital. A outra não. Esses episódios de má sorte são o que tornam a vida tão injusta. Mas, se há algum consolo, são eles que deixam a vida um pouco mais fácil de ser estudada pelos economistas. Eles usam as arbitrariedades da vida para buscar efeitos causais.

Dos 43 presidentes norte-americanos, 16 foram vítimas de tentativas de assassinato, e 4 foram mortos. Os motivos para alguns deles terem sobrevivido foram essencialmente aleatórios.

Vejamos John F. Kennedy e Ronald Reagan. Ambos foram alvo direto de balas nas partes mais vulneráveis do corpo. A bala que atingiu JFK explodiu seu cérebro, matando-o pouco tempo depois. A bala que atingiu Reagan parou a centímetros de seu coração, permitindo aos médicos salvarem sua vida. Reagan sobreviveu, JFK morreu, sem razão aparente — apenas sorte.

Esses atentados contra a vida de líderes e a arbitrariedade de seus resultados é algo que acontece em todo o mundo. Vejamos Akhmad Kadyrov, da Chechênia, e Adolf Hitler, da Alemanha. Ambos estiveram a centímetros da explosão de uma bomba. Kadyrov morreu. Hitler tinha mudado sua agenda, acabou deixando a sala com o explosivo alguns minutos mais cedo para pegar um trem, e por isso sobreviveu.

E podemos usar a fria aleatoriedade da vida — o assassinato de Kennedy, mas não de Reagan — para ver o que acontece, em média, quando o líder de uma nação é assassinado. Dois economistas, Benjamin F. Jones e Benjamin A. Olken, fizeram exatamente isso. O grupo de controle aqui é qualquer país nos anos imediatamente após um quase assassinato — por exemplo, os Estados Unidos em meados da década de 1980. O grupo de tratamento é qualquer país nos anos imediatamente após um assassinato — por exemplo, os Estados Unidos em meados da década de 1960.

Qual, então, é o efeito de ter seu líder assassinado? Jones e Olken descobriram que assassinatos bem-sucedidos alteram dramaticamente a história do mundo, levando países por caminhos radicalmente diferentes. Um novo líder faz com que países antes pacíficos vão a guerra e nações antes belicosas a alcançar a paz. Um novo líder faz com que economias prósperas entrem em recessão e nações em crise prosperem.

Na verdade, os resultados desse experimento natural baseado em assassinato subverteram algumas décadas de sabedoria convencional sobre como os países funcionam. Muitos economistas anteriormente tendiam acreditar que líderes eram em grande medida fantoches impotentes conduzidos por forças externas. Nem tanto, de acordo com a análise de Jones e Olken do experimento da natureza.

Muitos não considerariam essa análise de tentativas de assassinato de líderes mundiais um exemplo de Big Data. O número de líderes assassinados ou quase assassinados no estudo definitivamente foi pequeno — assim como o número de guerras que resultaram ou não deles. Os conjuntos de dados econômicos necessários para caracterizar a trajetória de uma economia eram grandes, mas em grande parte anteriores à digitalização.

Contudo, tais experimentos naturais — apesar de agora usados quase exclusivamente por economistas — são poderosos e assumirão uma crescente importância em uma era com mais, maiores e melhores conjuntos de dados. Essa é uma ferramenta que cientistas de dados tão cedo não irão abandonar.

E sim, como já deve estar claro agora, os economistas estão desempenhando um importante papel no desenvolvimento da ciência de dados. Pelo menos eu gostaria de pensar que sim, já que essa é minha formação.

Onde mais podemos encontrar experimentos naturais — em outras palavras, situações em que o curso aleatório de eventos coloque pessoas em grupos de tratamento e de controle?

O exemplo mais claro é a loteria, e é por isso que os economistas a adoram — não apostar, o que achamos irracional, mas estudá-la. Se uma bola de pingue-pongue com um três gravado cair do globo, o sr. Jones será um homem rico. Mas se for a bola com o número seis, será o sr. Johnson.

Para testar os efeitos causais de boladas inesperadas, os economistas comparam aqueles que ganham na loteria com os que compram o bilhete, mas perdem. Esses estudos em geral têm descoberto que ganhar na loteria não torna alguém feliz em curto prazo, mas sim em longo prazo.*

Economistas podem ainda utilizar a aleatoriedade das loterias para analisar como a vida de uma pessoa muda quando um vizinho fica rico. Os dados mostram que seu vizinho ganhar na loteria pode ter um impacto significativo também na sua vida. Por exemplo, se seu vizinho ganhar uma bolada, aumenta a probabilidade de que você compre um carro de luxo, como um BMW. Por quê? Os economistas defendem que é quase certo que a causa seja a inveja por seu vizinho rico ter comprado um carro de luxo. Atribua isso à natureza humana. Se o sr. Johnson vir o sr. Jones dirigindo um BMW novinho em folha, vai querer um também.

Infelizmente, o sr. Johnson em geral não pode bancar um carro assim, e é por isso que os economistas descobriram que vizinhos de ganhadores da loteria têm significativa maior probabilidade de se tornarem insolventes. Acompanhar o sr. Jones, neste caso, é impossível.

Mas os experimentos naturais não têm de ser explicitamente aleatórios, como as loterias. Uma vez que começa a procurar por aleatoriedade, passa a vê-la em todos os lugares — e você pode usá-la para entender como nosso mundo funciona.

Médicos fazem parte de um experimento natural. De vez em quando, o governo, por razões essencialmente arbitrárias, muda a fórmula que usa para reembolsar os médicos por pacientes do Medicare, serviço de assistência médica. Os médicos em alguns países veem seus honorários por determinados procedimentos subirem, enquanto caem em outros países.

Dois economistas — Jeffrey Clemens e Joshua Gottlieb, meu antigo colega de classe — testaram os efeitos dessa mudança arbitrária. Será que os médicos dispensam aos pacientes sempre os mesmos cuidados, o que eles entendem ser o mais necessário? Ou será que eles são levados por incentivos financeiros?

Os dados claramente mostram que os médicos podem ser motivados por incentivos monetários. Em países com os reembolsos mais altos, alguns médicos determinam substancialmente procedimentos mais bem-remunerados — mais cirurgias de catarata, colonoscopias e ressonâncias magnéticas, por exemplo.

E então, vem a grande pergunta: seus pacientes têm melhores resultados depois de receber todo esse cuidado extra? Clemens e Gottlieb relataram apenas “pequenos impactos na saúde”. Os autores não descobriram impacto estatisticamente significativo na mortalidade. Dê fortes incentivos financeiros para médicos determinarem determinados procedimentos, sugere esse experimento natural, e alguns pedirão mais procedimentos que não fazem grande diferença para a saúde dos pacientes e aparentemente não prolongam suas vidas.

Experimentos naturais podem ajudar a responder questões de vida ou morte. Eles também podem ajudar com aquelas que, para alguns jovens, parecem ser de vida ou morte.

Stuyvesant High School (conhecido como “Stuy”) fica em um prédio de dez andares de tijolos avermelhados de US$150 milhões de frente para o Rio Hudson, a algumas quadras do World Trade Center, na parte baixa de Manhattan. O Stuy é, em uma palavra, esplêndido. Ele oferece 55 cursos de Advanced Placement (AP) [aulas preparatórias rigorosas que concedem créditos extras nos testes de admissão das universidades], sete idiomas e disciplinas eletivas em história judaica, ficção científica e literatura ásio-americana. Aproximadamente um quarto dos formados são aceitos em universidades da Ivy League ou outras de igual prestígio. O Stuyvesant formou a professora de física de Harvard, Lisa Randall, o estrategista de Obama, David Axelrod, o ator ganhador do Oscar, Tim Robbins, e o romancista Gary Shteyngart. Entre seus paraninfos estão Bill Clinton, Kofi Annan e Conan O’Brien.

A única coisa mais impressionante do que os cursos e os graduados de Stuyvesant é seu custo: zero dólares. Ele é um colégio público e provavelmente o melhor do país. Na verdade, um estudo recente utilizou 27 milhões de opiniões de 300.000 alunos e pais para classificar todos os colégios de ensino médio dos Estados Unidos. O Stuy

ficou em primeiro lugar. Não é de se estranhar, então, que ambiciosos pais de classe média de Nova York e sua igualmente ambiciosa prole se tornaram obcecados pela marca da Stuy.

Para Ahmed Yilmaz*, filho de um agente de seguros e professor no Queens, Stuy era "*o* colégio de ensino médio".

"Famílias de imigrantes e da classe trabalhadora veem o Stuy como uma saída", explica Yilmaz. "Se seu filho frequenta o Stuy, ele vai para uma das melhores universidades. A família ficará bem."

Mas o que é preciso para entrar no Stuyvesant High School? Você precisa morar em um dos cinco distritos da cidade de Nova York e ter uma pontuação mínima na prova de admissão. Só isso. Não é preciso recomendações, trabalhos extras, ser filho de egressos, nem qualquer outra ação afirmativa. Um dia, uma prova, uma nota. Se sua nota estiver acima do mínimo, você entra.

Todo mês de novembro, aproximadamente 27.000 jovens novaiorquinos se submetem à prova de admissão. A concorrência é brutal. Menos de 5% daqueles que fazem a prova conseguem entrar em Stuy.

Yilmaz explica que sua mãe teve que "ralar muito" e usar centavos que conseguia juntar na preparação de seu filho para a prova. Depois de meses passando todas as tardes dos dias de semana e finais de semana inteiros se preparando, Yilmaz estava confiante de que entraria em Stuy. Ele ainda se lembra do dia que recebeu o envelope com o resultado. Ele não conseguiu por duas questões.

Perguntei a ele o que sentiu. "Como se sentiria", respondeu ele, "se seu mundo inteiro desmoronasse quando você estava no ensino médio?"

Seu prêmio de consolação não era nem de longe uma tragédia — Bronx Science, outra escola pública exclusiva e muito bem classificada. Mas não era Stuy. E Yilmaz achava que o Bronx Science era uma escola especializada destinada a áreas mais técnicas. Quatro anos depois, ele foi rejeitado em Princeton. Ele frequentou a Tufts e mudou algumas vezes de carreira. Hoje ele é um funcionário razoavelmente bem-sucedido de uma empresa de tecnologia, embora diga que seu trabalho é "tedioso" e nem tão bem remunerado como ele gostaria.

Mais de uma década depois, Yilmaz admite que às vezes imagina como a vida teria sido se ele tivesse entrado em Stuy. "Tudo seria diferente", diz ele. "Todo mundo que conheço seria diferente." Ele imagina se o Stuyvesant High School o teria levado a uma nota melhor no SAT [equivalente ao ENEM] e a uma universidade como Princeton ou Harvard (que ele considera significativamente melhores do que Tufts), e talvez a um emprego mais lucrativo ou gratificante.

Conjecturar hipóteses pode ser tanto divertido quanto torturante para os seres humanos. O que teria sido minha vida se eu tivesse me declarado para aquela garota ou aquele garoto? E se eu tivesse aceitado aquele emprego? Se eu tivesse ido para aquela escola? Mas esses "e se" parecem impossíveis de se responder. A vida não é um videogame. Não dá para revivê-la em diferentes cenários até que consiga os resultados que deseja.

Milan Kundera, escritor de origem tcheca, tem uma citação concisa sobre o tema em seu livro *A Insustentável Leveza do Ser*: "A vida humana acontece apenas uma vez, e o motivo pelo qual não podemos determinar quais de nossas decisões são boas e quais são ruins é que em determinada situação só podemos tomar uma decisão; não temos direito a uma segunda, terceira ou quarta vida para comparar nossas decisões."

Yilmaz nunca vivenciará uma vida em que ele tenha conseguido dois pontos extras naquela prova.

Mas talvez exista uma maneira de podermos ter alguma ideia do quão diferente poderia ou não ter sido sua vida estudando um grande número de alunos do Stuyvesant High School.

A metodologia direta e ingênua seria comparar todos os alunos que foram a Stuyvesant e todos os que não foram. Poderíamos analisar seu desempenho nas provas de AP e de SAT — e em quais universidade eles foram aceitos. Com isso, teríamos descoberto que os alunos que frequentaram o Stuyvesant obtiveram notas muito mais altas em testes padronizados e foram aceitos em universidades substancialmente melhores. Mas, como já vimos neste capítulo, esse tipo de evidência, por si só, não é convincente. Talvez a razão pela qual os alunos de Stuyvesant tenham se saído bem melhor seja o fato do Stuy atrair melhores alunos. A correlação aqui prova causalidade.

Para testar os efeitos causais do Stuyvesant High School, precisamos comparar dois grupos que sejam quase idênticos: um de tratamento, que frequentou Stuy, e outro que não frequentou, o grupo de controle. Precisamos de um experimento natural. Mas onde podemos encontrá-lo?

A resposta: alunos, como Yilmaz, que obtiveram notas muito, muito próximas a mínima necessária para entrar no Stuyvesant.* Alunos que quase entraram são o grupo de controle; alunos que obtiveram a nota mínima são o grupo de tratamento.

Há poucas razões para suspeitar que os alunos nas duas faixas limítrofes da nota de corte sejam muito diferentes em talento ou determinação. O que, afinal de contas, faz com que uma pessoa tenha uma nota um ou dois pontos mais alta do que outra? Talvez aquela com nota um pouco mais baixa tenha dormido dez minutos a menos ou tomado um café da manhã menos nutritivo. Talvez aquela que tenha nota mais alta tenha se lembrado de uma palavra especialmente difícil na prova por causa de uma conversa que teve com a avó três anos antes.

Na verdade, essa categoria de experimentos naturais — utilizando notas de corte precisas — é tão poderosa que tem seu próprio nome entre os economistas: regressão descontinuada. Toda vez que há um número preciso que divida pessoas em dois grupos distintos — uma descontinuidade — os economistas podem comparar — ou fazer a análise regressiva — os resultados das pessoas bem próximos à nota de corte.

Dois economistas, M. Keith Chen e Jesse Shapiro, utilizaram o corte preciso usado pelas prisões federais para testar os efeitos das condições severas das prisões em crimes futuros. Prisioneiros federais nos Estados Unidos recebem uma classificação com base na natureza de seus crimes e em seu histórico criminal. Esse número determina as condições de encarceramento. Aqueles com "notas" altas o bastante vão para um presídio de segurança máxima, o que significa menos contato com outras pessoas, menos liberdade de movimento e provavelmente mais violência dos carcereiros e dos outros prisioneiros.

Novamente, não seria justo comparar todo o universo de prisioneiros em um estabelecimento correcional de segurança máxima com todo o universo de prisioneiros em prisões de segurança mínima. As prisões de segurança máxima conterão mais assassinos e estupradores, enquanto as prisões de segurança mínima conterão mais traficantes e ladrões comuns.

Mas aqueles logo acima ou abaixo do patamar de corte têm históricos criminais praticamente idênticos. Essa minúscula diferença, porém, significou uma experiência muito diferente na prisão.

O resultado? Os economistas descobriram que os prisioneiros enviados para condições mais severas tinham maior probabilidade de cometer novos crimes depois de serem libertados. As prisões com condições mais severas, em vez de impedir a reincidência, tornaram os presos mais endurecidos e mais violentos quando retornaram para o mundo exterior.

Então, o que mostrou a "regressão descontinuada" para o Stuyvesant High School? Uma equipe de economistas do MIT e da Universidade de Duke — Atila Abdulkadiroĝlu, Joshua Angrist e Parag Pathak — conduziram o estudo. Eles compararam os resultados dos pupilos de Nova York nos dois lados da nota de corte. Em outras palavras, esses economistas analisaram centenas de alunos que, assim como Yilmaz, *não ingressaram* em Stuyvesant por uma questão ou duas. Eles os compararam com centenas de alunos que se saíram melhor na prova e *ingressaram* em Stuy por uma ou duas questões. Suas medidas de sucesso foram as notas no AP, no SAT e as classificações das universidades que cada um deles frequentou.

Os impressionantes resultados obtidos se evidenciaram pelo título atribuído ao artigo: "Ilusão da Elite". Os efeitos do Stuyvesant High School? Nulos. Nada. Zero. Os alunos nos dois lados da nota de corte terminaram com escores de AP e de SAT indistinguíveis e frequentaram universidades de igual prestígio.

Os pesquisadores concluíram que a única razão de os alunos de Stuy terem alcançado maior sucesso na vida do que os não alunos de Stuy é que aqueles que ingressam em Stuyvesant são alunos melhores. O Stuy não é a *causa* de seu melhor desempenho nas provas de AP ou SAT, ou de terem frequentado melhores universidades.

"A concorrência intensa nas provas para vagas no colégio", escreveram os economistas, " não parece se justificar por um melhor aprendizado para uma ampla gama de alunos."

Por que a escola que você frequenta não importa? Mais algumas histórias podem nos ajudar na resposta. Vejamos outras duas alunas, Sarah Kaufmann e Jessica Eng, duas jovens novaiorquinas que sonhavam desde muito crianças em ir para Stuy. A nota de Kaufmann foi exatamente a de corte; ela conseguiu entrar por uma questão. "Não acho que outra coisa será mais emocionante do que isso", recorda Kaufmann. A nota de Eng foi logo abaixo do corte; ela perdeu a vaga por uma questão. Kaufmann ingressou na escola de seus sonhos, Stuy. Eng, não.

Então, como foi a vida delas? Ambas tiveram carreiras bem-sucedidas e gratificantes — assim como a maioria das pessoas com notas entre os 5% dos novaiorquinos melhores qualificados nas provas. Eng, ironicamente, gostou mais de sua experiência no ensino médio. O Bronx Science, colégio que frequentou, era o único com um museu do Holocausto. Eng descobriu que adorava curadoria e estudou antropologia em Cornell.

Kaufmann se sentiu um pouco perdida em Stuy, onde os alunos eram muito focados em notas e ela achou que a ênfase maior era nas provas e não em ensinar. Ela chamou sua experiência de "definitivamente uma confusão". Mas foi um aprendizado. Ela percebeu que só se candidataria para universidades de ciências humanas, que davam mais ênfase ao aprendizado. Ela foi aceita em sua faculdade dos sonhos, a Universidade de Wesleyan. Lá ela encontrou a paixão em ajudar os outros e agora trabalha como defensora pública.

As pessoas se adaptam a sua experiência, e aquelas que terão sucesso encontram vantagens em qualquer situação. Os fatores que tornam alguém bem-sucedido são o talento e a determinação. E não paraninfos prestigiados ou outras vantagens que uma escola de renome possa oferecer.

Esse é apenas um estudo, e provavelmente seus resultados são enfraquecidos pelo fato de a maioria dos alunos que quase entraram no Stuyvesant acabaram em outra excelente escola. Mas há crescentes evidências de que, embora ir para uma boa escola seja importante, há pouca vantagem em frequentar a melhor escola possível.

Vejamos a faculdade. Será que faz diferença frequentar uma das melhores universidades do mundo, como Harvard, ou uma boa faculdade como a Universidade Estadual da Pensilvânia?

Mais uma vez, há uma clara correlação entre a classificação da universidade em que uma pessoa se forma e sua renda. Aos dez anos de carreira, um graduado médio de Harvard ganha US$123 mil por ano. Enquanto que um graduado na Universidade Estadual da Pensilvânia ganha US$87.800.

Mas essa correlação não implica em causalidade.

Dois economistas, Stacy Dale e Alan B. Krueger, pensaram em uma maneira engenhosa de testar o papel causal das universidades de elite no potencial rendimento futuro de seus graduados. Eles tinham um grande conjunto de dados que rastreou toda uma fonte de informações sobre alunos do ensino médio, incluindo quais faculdades eles se candidataram e em quais eles foram aceitos, históricos familiares e suas rendas quando adultos.

Para obter um grupo de tratamento e um de controle, Dale e Krueger compararam alunos com históricos semelhantes que foram aceitos pelas mesmas universidades, mas que escolheram diferentes. Alguns alunos que foram aceitos em Harvard, mas que frequentaram a Universidade Estadual da Pensilvânia — talvez para ficarem mais perto de namoradas ou namorados ou porque havia um professor específico com quem queriam estudar. Esses estudantes, em outras palavras, eram tão talentosos, de acordo com os conselhos de admissão, quanto aqueles que frequentaram Harvard. Mas tiveram experiências educacionais diferentes.

Então, quando dois alunos, de históricos semelhantes, entraram em Harvard, mas um deles escolheu ir para Universidade Estadual da Pensilvânia, o que aconteceu? Os resultados dos pesquisadores foram tão surpreendentes quanto os do estudo do Stuyvesant High School. Eles terminaram com mais ou menos a mesma renda anual em suas carreiras. Se o futuro salário é a medida, alunos parecidos aceitos em faculdades de prestígio semelhantes que escolhem frequentar escolas diferentes acabam chegando praticamente no mesmo patamar.

Nossos jornais repletos de artigos sobre pessoas de gigantesco sucesso que frequentaram universidades da Ivy League: pessoas como o fundador da Microsoft, Bill Gates, e os fundadores do Facebook, Mark Zuckerberg e Dustin Moskovitz, todos alunos de Harvard. (Todos abandaram a universidade, o que levanta mais dúvidas sobre o valor de uma educação no padrão Ivy League.)

Há ainda muitas histórias de pessoas que eram talentosas o bastante para serem aceitas em uma universidade da Ivy League, que escolheram uma escola menos prestigiosa, e tiveram vidas incrivelmente bem-sucedidas: pessoas como Warren Buffett, que entrou para a Wharton School na Universidade da Pensilvânia, uma faculdade de administração da Ivy League, mas que se transferiu para a Universidade de Nebraska–Lincoln, porque era mais barata, por odiar a Filadélfia e achar que as aulas em Wharton eram chatas. Os dados sugerem que, pelos menos do ponto de vista dos rendimentos, escolher frequentar uma universidade de menor prestígio é uma boa decisão, pelo menos para Buffett e outros.

Este livro se chama *Todo Mundo Mente*. Com isso, quero dizer principalmente que as pessoas mentem — para os amigos, para as pesquisas e para si mesmas — para transmitirem uma melhor imagem.

Mas o mundo também mente para nós ao nos apresentar dados equivocados e ilusórios. O mundo nos mostra um gigantesco número de bem-sucedidos egressos de Harvard, mas menos graduados de sucesso da Universidade Estadual da Pensilvânia, e nós presumimos que há uma enorme vantagem em frequentar Harvard.

Ao extrair com sabedoria o significado dos experimentos naturais, podemos compreender corretamente os dados do mundo — para descobrir o que realmente é útil e o que não é.

Experimentos naturais também dizem respeito ao capítulo anterior. Eles frequentemente requerem ajustar o foco — nos grupos de tratamento e de controle: as cidades no experimento do Super Bowl, os países no experimento dos preços do Medicare, os estudantes com nota próxima ao corte no experimento do Stuyvesant. E focar, como discutido no capítulo anterior, geralmente requer conjuntos de dados grandes e abrangentes — do tipo que cada vez mais estão disponíveis conforme o mundo se torna digital. Como não sabemos quando a natureza escolherá conduzir seus experimentos, não podemos montar um pequeno estudo para medir seus resultados. Precisamos de muitos dados existentes para aprender com essas intervenções. Precisamos de Big Data.

Há mais uma questão importante sobre os experimentos — os nossos próprios ou os da natureza — detalhados neste capítulo. Grande parte deste livro foca em compreender o mundo — qual o custo do racismo para Obama, quantos homens são realmente gays, como homens e mulheres são inseguros sobre o próprio corpo. Mas esses experimentos controlados ou naturais têm uma aptidão mais prática. Eles pretendem melhorar nossa tomada de decisão, para nos ajudar a saber quais intervenções funcionam e quais, não.

A empresas são capazes de aprender como obter mais clientes. O governo consegue descobrir como usar o reembolso para motivar mais os médicos. Alunos podem saber quais escolas terão mais valor. Esses experimentos demonstram o potencial do Big Data para substituir os palpites, a sabedoria convencional e correlações fajutas por o que realmente funciona — *causalidade*.

* Field goal é uma forma de pontuar no futebol americano. Consiste em chutar a bola oval de qualquer ponto do campo entre os postes de gol. Vale 3 pontos no placar.

* Um famoso artigo de 1978 que alegava que ganhar na loteria não torna alguém feliz foi totalmente desacreditado.

* Mudei o nome e alguns detalhes.

* Ao buscar por pessoas como Yilmaz, que tiveram nota muito próxima do corte, fiquei impressionado pelo número de pessoas — na faixa de 20 aos 50 anos — que se lembraram de sua experiência no dia da prova e falaram em termos bastante dramáticos sobre não ter conseguido entrar. Entre eles está o ex-congressista e candidato a prefeito da cidade de Nova York Anthony Weiner, que diz ter deixado de ingressar no Stuy por um único ponto. “Eles não me quiseram”, disse-me em nossa entrevista por telefone.

PARTE III

BIG DATA: USE COM CUIDADO

7

BIG DATA É CASCATA? O QUE ELE NÃO É CAPAZ DE FAZER

“ **S** eth, Lawrence Summers gostaria de conhecê-lo”, dizia o e-mail, de um modo um tanto enigmático. Era de um de meus orientadores de doutorado, Lawrence Katz. Katz não me dizia por que Summers estava interessado em meu trabalho, mas depois descobri que Katz o acompanhava desde o início.

Sentei na sala de espera do escritório de Summers. Depois de um pequeno atraso, o ex-secretário do Tesouro dos Estados Unidos, ex-reitor de Harvard e vencedor de alguns dos maiores prêmios em economia, me convidou a entrar.

Summers começou a reunião lendo meu artigo, que seu secretário imprimira para ele, sobre como o racismo afetou Obama. Summers lê muito rápido. Enquanto lia, ocasionalmente punha a língua para fora para a direita, enquanto seus olhos rapidamente moviam-se da esquerda para a direita descendo a página. Summers lendo um artigo sobre ciência social lembrava um grande pianista executando uma sonata. Ele estava tão concentrado que parecia esquecer todo o resto a sua volta. Em menos de cinco minutos concluiu a leitura de meu artigo de trinta páginas.

“Você diz que as buscas no Google pelo termo ‘nigger’ sugerem racismo”, disse Summers. “Isto parece plausível. Elas preveem onde Obama recebe menos apoio do que Kerry. É interessante. Podemos realmente pensar em Obama e Kerry como iguais?”

“Eles foram classificados como tendo ideologias semelhantes por cientistas políticos”, respondi. “Também, não há correlação entre o racismo e as mudanças na votação da Câmara dos Representantes. O resultado continua substancialmente equilibrado mesmo quando acrescentamos controles demográficos, frequência à igreja e posse de arma.” É assim que nós, economistas, falamos. Minha animação aumentava.

Summers parou e me encarou. Ele olhou brevemente para a TV em seu escritório, sintonizada na CNBC, depois me encarou novamente, olhou de novo para a TV e novamente para mim. “Certo, gostei deste artigo”, disse Summers. “Em que mais você está trabalhando?”

Os sessenta minutos seguintes devem ter sido os mais intelectualmente estimulantes de minha vida. Summers e eu conversamos sobre taxas de juros e inflação, policiamento e crimes, negócios e caridade. Existe uma razão para que tantas pessoas que conhecem Summers fiquem fascinadas. Tive a sorte de conversar com pessoas incrivelmente inteligentes em minha vida; Summers me impressionou como o mais inteligente. Ele é obcecado por ideias, mais do que tudo, o que parece frequentemente colocá-lo em apuros. Ele renunciara a reitoria de Harvard depois de sugerir a possibilidade de que parte do motivo para a escassez de mulheres na ciência é o fato de homens terem uma variação maior em seus QIs. Se ele acha uma ideia interessante, a tendência de Summers é dizê-la, mesmo que possa ofender algumas pessoas.

Já havíamos passado meia hora do horário programado para o fim de nossa reunião. A conversa estava deliciosa, mas eu ainda não fazia a menor ideia do porquê estava ali, nem quando deveria ir embora, nem como saberia que era hora de partir. Tive a sensação, àquela altura, de que o próprio Summers tinha esquecido por que marcara a reunião.

E então ele fez a pergunta de um milhão — ou talvez bilhão. “Você acha que pode prever o mercado de ações com estes dados?”

Ah. Finalmente o motivo de Summers ter me chamado a seu escritório.

Summers não é nem de longe a primeira pessoa a me perguntar isto. Meu pai em geral apoia meus interesses não convencionais de pesquisa. Mas certa vez ele resolveu abordar o assunto. “Racismo, abuso infantil, aborto”,

disse ele. “Não dá para você ganhar algum dinheiro com essa sua especialidade?” Amigos e outros familiares também já levantaram a questão. Assim como colegas de trabalho e estranhos na internet. Todo mundo parece querer saber se posso utilizar as buscas no Google — ou outro Big Data — para escolher ações. Agora foi o ex-secretário do Tesouro dos Estados Unidos. Isto era bem mais sério.

Então, as novas fontes de Big Data *podem* prever com sucesso quais as tendências de movimentação de ações? A resposta curta é não.

Nos capítulos anteriores discutimos os quatro poderes do Big Data. Este capítulo é sobre suas limitações — tanto o que não é possível fazer com ele e, oportunamente, o que não devemos fazer. E um lugar para começar é contando a história da tentativa fracassada, minha e de Summers, de vencer o mercado.

No Capítulo 3, observamos que os novos dados têm maior probabilidade de render grandes resultados quando a pesquisa existente em determinado campo é fraca. Infelizmente é verdade que em nosso mundo de hoje é muito mais fácil obter novas perspectivas sobre racismo, abuso infantil ou aborto do que percepções novas e lucrativas sobre o desempenho de um negócio. Isto porque recursos gigantescos já são destinados para procurar por qualquer mínima vantagem na medição do desempenho de um negócio. A concorrência na área de finanças é feroz. Isso já é um ponto a menos para nós.

Summers, que não é conhecido exatamente por ser efusivo sobre a inteligência de outras pessoas, estava certo de que os fundos hedge já estavam muito mais adiantados do que nós. Fiquei impressionado, durante nossa conversa, com quanto respeito ele tinha por eles e quantas de minhas sugestões ele estava convencido de que já haviam sido pensadas. Orgulhosamente falei do algoritmo que havia criado que me permitia obter dados mais completos do Google Trends. Ele disse que era inteligente. Quando perguntei a ele se o Renaissance, um fundo hedge quantitativo, teria imaginado esse algoritmo, ele riu e disse: “Sim, claro que teriam descoberto.”

A dificuldade de acompanhar os fundos hedge não era o único problema fundamental que Summers e eu tínhamos que superar ao utilizar grandes e novos conjuntos de dados para vencer o mercado de ações.

A MALDIÇÃO DA DIMENSIONALIDADE

Suponha que sua estratégia para prever o mercado de ações seja encontrar a moeda da sorte — mas uma que será descoberta através de cuidadosos testes. Esta é sua metodologia: você identifica mil moedas — de 1 a mil. Toda manhã, por dois anos, joga a moeda para cima, registra se o resultado foi cara ou coroa, e depois anota se o Índice S&P subiu ou desceu naquele dia. Analisa todos seus dados e *voilà*! Você descobre algo. Ocorre que 70,3% das vezes que a Moeda 391 deu cara, o Índice S&P subiu. A relação é alta e estatisticamente significativa. Você acaba de encontrar sua moeda da sorte!

Basta jogar a Moeda 391 toda manhã e comprar ações sempre que der cara. Seus dias de camisetas baratas e macarrão instantâneo no jantar acabaram. A Moeda 391 é seu ingresso para a boa vida!

Ou não.

Você se tornou mais uma vítima de um dos mais diabólicos aspectos da “maldição da dimensionalidade”. Ela pode surgir sempre que existem muitas variáveis (ou “dimensões”) — neste caso, mil moedas — em busca de poucas observações — no exemplo, 504 dias de negociação ao longo de dois anos. Uma dessas dimensões — a Moeda 391, neste caso — provavelmente terá sorte. Diminua o número de variáveis — jogue apenas cem moedas — e haverá bem menos probabilidade de uma delas ter sorte. Aumente o número de observações — tente prever o comportamento do Índice S&P por 20 anos — e as moedas terão dificuldade de acompanhar.

A maldição da dimensionalidade é um grande problema do Big Data, como os conjuntos de dados mais novos exponencial e frequentemente nos oferecem mais variáveis do que as fontes de dados tradicionais — cada termo de busca, cada categoria de tuíte etc. Muitas pessoas que alegam prever o mercado de ações utilizando alguma fonte de Big Data não passam de vítimas da maldição. Tudo que fizeram, na verdade, foi encontrar o equivalente à Moeda 391.

Vejamos, por exemplo, uma equipe de cientistas de computação da Universidade de Indiana e da Universidade de Manchester que alegaram ser capazes de prever qual a movimentação do mercado de ações com base no que as pessoas estavam tuitando. Eles criaram um algoritmo para codificar os humores diários do mundo com base em tuítes. Usaram técnicas semelhantes à análise de estado emocional discutida no Capítulo 3. Entretanto, não codificaram apenas um estado emocional, mas muitos — felicidade, raiva, gentileza e mais. Eles descobriram que a preponderância de tuítes sugerindo tranquilidade, como “Estou calmo”, prevê que o Dow Jones Industrial Average suba seis dias depois. Um fundo hedge foi criado para explorar essas descobertas.

Qual é o problema aqui?

A questão fundamental é que eles testaram coisas demais. E se testarmos coisas o bastante, apenas pela aleatoriedade, uma delas será estatisticamente significativa. Eles testaram emoções demais. E testaram cada emoção um dia, dois, três, até sete dias antes do comportamento do mercado de ações que tentavam prever. E todas essas variáveis foram usadas para explicar apenas alguns meses das oscilações do Dow Jones.

A tranquilidade seis dias antes não era um preditor legítimo do mercado de ações. Seis dias antes eram o equivalente do Big Data de nossa hipotética Moeda 391. O fundo hedge baseado em tuítes foi encerrado um mês depois de ser criado em razão dos resultados medíocres.

Fundos hedge que tentam avaliar o mercado com tuítes não são os únicos que enfrentam a maldição da dimensionalidade. Assim como os inúmeros cientistas que tentaram descobrir os segredos da genética de quem somos.

Graças ao Projeto do Genoma Humano, agora é possível coletar e analisar o DNA humano inteiro. O potencial deste projeto é gigantesco.

Talvez descubramos o gene que causa a esquizofrenia. Ou o que provoca o Mal de Alzheimer e o de Parkinson ou a Esclerose Lateral Amiotrófica. Talvez possamos descobrir o gene da inteligência. Existe um gene que acrescenta muitos pontos no QI? Haveria um gene responsável pela genialidade?

Em 1998, Robert Plomin, um proeminente geneticista comportamental, alegou ter descoberto a resposta. Ele recebeu um conjunto de dados que incluiu o DNA e o QI de centenas de estudantes. Ele comparou o DNA de “gênios” — aqueles com QI de 160 ou superior — com o daqueles com QIs medianos.

Ele descobriu uma impressionante diferença no DNA desses dois grupos. Localizava-se em um pequeno recanto do cromossomo 6, um gene obscuro, mas poderoso, usado no metabolismo do cérebro. Uma versão deste gene, chamada IGF2r, era duas vezes mais comum em gênios.

“O Primeiro Gene a Ser Relacionado com a Alta Inteligência Foi Descoberto”, dizia a manchete do *New York Times*.

É possível pensar em muitas questões éticas decorrentes da descoberta de Plomin. Devemos permitir que os pais examinem os filhos em busca do gene IGF2r? Devemos permitir que os pais abortem um feto com uma variante de baixo QI? Devemos modificar geneticamente as pessoas para que tenham um QI mais alto? O gene IGF2r se correlaciona com raça? Queremos saber a resposta para esta pergunta? A pesquisa sobre a genética do QI deve continuar?

Antes que os bioéticos tenham que enfrentar estas questões espinhosas, havia uma pergunta mais básica para os geneticistas, incluindo o próprio Plomin. O resultado foi preciso? Era mesmo verdade que o gene IGF2r é capaz de prever o QI? Que os gênios têm duas vezes mais probabilidade de apresentar uma determinada variante deste gene?

Não. Alguns anos depois do seu estudo original, Plomin teve acesso a outra amostra de pessoas, que também incluía o DNA e a pontuação de QI. Desta vez, o gene IGF2r não se correlacionou com o QI. Plomin — e isto é um sinal de um bom cientista — se retratou da alegação.

Isto, na verdade, tem sido um padrão geral na pesquisa envolvendo genética e QI. Primeiro, os cientistas relatam ter descoberto uma variante genética que prevê o QI. Depois, eles obtêm novos dados e descobrem que a afirmação original estava errada.

Por exemplo, em um artigo recente, uma equipe de cientistas, liderada por Christopher Chabris, examinou doze importantes alegações sobre as variantes genéticas associadas ao QI. Analisaram os dados de dez mil pessoas. E não conseguiram reproduzir a correlação para qualquer das doze alegações.

Qual é o problema com todas essas alegações? A maldição da dimensionalidade. O genoma humano, os cientistas agora sabem, difere em milhões de maneiras. Existem, simplesmente, genes demais para avaliar.

Se testarmos um grande número de tuítes para verificar se há correlação com o mercado de ações, descobriremos uma correlação apenas pelo acaso. Se testarmos variantes genéticas suficientes para verificar se há alguma correlação com o QI, encontraremos uma que se correlacione simplesmente pelo acaso.

Como superar a maldição da dimensionalidade? É preciso ter um pouco de humildade sobre seu trabalho e não se apaixonar pelos resultados. Você tem que submeter os resultados a testes adicionais. Por exemplo, antes de apostar todas suas economias na Moeda 391, vai querer saber como se sai nos próximos anos. Cientistas sociais chamam isto de teste “fora da amostra”. E quanto mais variáveis você testa, mais humilde tem que ser. Quanto mais variáveis testa, mais difícil tem que ser o teste fora da amostra. É também crucial registrar todos os testes que realizar. Então, é possível saber exatamente qual é a probabilidade de se tornar vítima da maldição e o quanto precisa ser cético com seus resultados. O que nos leva de volta a Larry Summers e eu. Veja como tentamos vencer o mercado de ações.

A primeira ideia de Summers foi usar as buscas para prever as vendas futuras de produtos-chave, como iPhones, que oferecem uma nova perspectiva no futuro desempenho da ação de uma companhia, como a Apple. Havia, de fato, uma correlação entre buscas por “iPhones” e suas vendas. Quando as pessoas pesquisam muito no

Google por “iPhones”, pode apostar que muitos celulares estão sendo vendidos. Entretanto, essa informação já estava incorporada no preço da ação da Apple. Claramente, quando havia muitas buscas no Google por “iPhones”, os fundos hedge também já haviam percebido que seria um grande sucesso de vendas, independentemente de terem usado os dados de busca ou outra fonte.

A próxima ideia de Summers foi prever futuros investimentos em países em desenvolvimento. Se um grande número de investidores pretende investir em países como o Brasil ou o México no futuro próximo, as ações das companhias nestes países com certeza irão subir. Talvez poderíamos prever um aumento no investimento com buscas-chave no Google — como “investir no México” ou “oportunidades de investimento no Brasil”. Isto se mostrou um beco sem saída. O problema? As buscas eram muito raras. Em vez de revelar padrões significativos, estes dados de busca se mostraram totalmente inconsistentes.

Tentamos buscas por ações individuais. Talvez se as pessoas pesquisassem por “GOOG”, isso significasse que pretendiam comprar ações da Google. Essas buscas pareciam predizer se as ações seriam muito negociadas. Mas não previam se as ações iriam subir ou cair. Uma grande limitação é que essas buscas não nos disseram se alguém estava interessado em comprar ou vender a ação.

Um dia, mostrei a Summers, com entusiasmo, uma nova ideia: buscas passadas por “comprar ouro” pareciam ter correlação com aumentos futuros no preço do ouro. Summers me disse que eu deveria testar a ideia no futuro para verificar se continuava precisa. Ela parou de funcionar, talvez porque alguns fundos hedge tenham descoberto a mesma relação.

No final, depois de alguns meses, não encontramos nada útil para nossos testes. Sem dúvida, se tivéssemos procurado por uma correlação com o desempenho do mercado para cada um dos bilhões de termos pesquisados no Google, teríamos encontrado uma que funcionaria, ainda que de modo frágil. Mas provavelmente seria apenas nossa Moeda 391.

A ÊNFASE EXCESSIVA NAQUILO QUE É MENSURÁVEL

Em março de 2012, Zoë Chance, professora de marketing da Universidade de Yale, recebeu um pequeno pedômetro branco pelo correio em seu escritório no centro de New Haven, Connecticut. Ela pretendia estudar como este dispositivo, que mede os passos dados durante o dia e lhe atribui pontos, inspira você a se exercitar mais.

O que aconteceu em seguida, conforme seu relato em uma palestra na TEDx, foi um pesadelo do Big Data. Chance se tornou tão obcecada e viciada em aumentar seus pontos que começou a andar a toda hora, da cozinha para a sala, para a sala de jantar, para o porão, em seu escritório. Ela caminhava pela manhã, tarde da noite, quase todas as horas do dia — 20 mil passos em um período de 24 horas. Ela checava o pedômetro centenas de vezes por dia, e grande parte do que restou de sua interação humana era com outros usuários do pedômetro através da internet, discutindo estratégias de como aumentar sua pontuação. Ela se recorda de colocar o pedômetro na filha de três anos toda vez que a menina andava, pois estava absolutamente obcecada por aumentar sua pontuação.

Chance ficou tão obstinada por aumentar seus números que perdeu completamente a noção. Ela se esqueceu da razão pela qual uma pessoa gostaria de aumentar a pontuação — exercitar-se, e não colocar a filha para contabilizar passos. Ela sequer concluiu a pesquisa acadêmica sobre o pedômetro. Finalmente, se livrou do aparelho depois de ir dormir muito tarde, exausta, por tentar ganhar mais alguns passos. Embora seja uma pesquisadora orientada para dados por profissão, a experiência a afetou profundamente. “Isso me tornou um tanto cética sobre se obter acesso a mais dados é sempre uma coisa boa”, diz Chance.

Esta é uma história extrema; mas indica um potencial problema de as pessoas usarem dados para tomar decisões. Números são sedutores. Podemos acabar obcecados por eles e, com isso, perdermos de vista considerações mais importantes. Zoë Chance se esqueceu, de certa forma, do resto de sua vida.

Mesmo paixões menos obsessivas por números apresentam desvantagens. Vejamos a ênfase moderna dos testes das escolas norte-americanas — julgando os professores com base nas notas dos alunos. Embora o desejo por medidas mais objetivas do que acontece nas salas de aulas seja legítimo, há muitas coisas que não podem ser prontamente captadas pelos números. Além disso, todos esses testes pressionaram muitos professores a ensinar focando apenas os testes — e pior. Um pequeno número, conforme demonstrado em um artigo de Brian Jacob e Steven Levitt, trapaceou abertamente na administração dos testes.

O problema é o seguinte: as coisas que podemos medir frequentemente não são exatamente aquelas com que nos importamos. Somos capazes de medir como os alunos se saem em questões de múltipla escolha. Podemos facilmente medir o pensamento crítico, a curiosidade ou o desenvolvimento pessoal. Apenas tentar aumentar um único número fácil de medir — testar notas ou os números de passos dados em um dia — nem sempre ajuda a alcançar o objetivo pretendido.

Em seus esforços para melhorar seu site, o Facebook também corre este risco. A companhia tem toneladas de dados das pessoas que utilizam o site. É fácil verificar se uma história do feed de notícias foi curtida, clicada, comentada ou compartilhada. Mas, de acordo com Alex Peysakhovich, cientista de dados do Facebook, com quem escrevo sobre esses assuntos, nenhum desses é um substituto perfeito para questões mais importantes: Qual foi a experiência ao utilizar o site? A história conectou o usuário com seus amigos? Ele trouxe novas informações do mundo? Ele se divertiu?

Ou vejamos a revolução dos dados do beisebol na década de 1990. Muitas equipes começaram a usar estatísticas intrincadas — em vez de contar com os velhos observadores humanos — para tomar decisões. Era fácil medir ataques e arremessos, mas não defesas, então algumas organizações acabaram subestimando a importância da defesa. Na verdade, em seu livro, *O Sinal e o Ruído,* Nate Silver estima que a Oakland A's, organização direcionada a dados representada em *O Homem que Mudou o Jogo,* abria mão de oito a dez vitórias por ano em meados da década de 1990 por causa de sua terrível defesa.

A solução nem sempre é mais o Big Data. Em geral, precisamos de um toque especial para o ajudar a funcionar melhor: o julgamento de pessoas e as pequenas pesquisas, o que podemos chamar de Small Data. Em uma entrevista com Silver, Billy Beane, então gerente geral da A's e o personagem principal de *O Homem que Mudou o Jogo,* disse que, na verdade, havia começado a aumentar o orçamento para observadores.

Para preencher as lacunas nesta gigantesca piscina de dados, o Facebook também teve que conduzir uma abordagem mais tradicional: perguntar o que as pessoas pensavam. Todo dia, ao carregar o feed de notícias, centenas de usuários do Facebook recebem perguntas sobre as histórias que aparecem em suas telas. Os conjuntos de dados do Facebook automaticamente coletados (curtidas, cliques e comentários) são suplementados, em outras palavras, por dados menores ("Você quer ver este post novamente em seu feed de notícias?", "Por quê?"). Sim, mesmo uma organização incrivelmente bem-sucedida na utilização de Big Data como o Facebook às vezes utiliza a fonte de informação mais desprezada neste livro: a pequena pesquisa.

Na verdade, por causa da necessidade de Small Data como complemento para sua ação principal — gigantescas coleções de cliques, curtidas e posts —, as equipes de dados do Facebook são um tanto diferentes do que você possa imaginar. O Facebook emprega psicólogos sociais, antropólogos e sociólogos exatamente para descobrir o que os números ignoram.

Alguns educadores, também, estão cada vez mais alertas aos pontos cegos no Big Data. Existe um crescente esforço nacional para suplementar os testes em massa por Small Data. As pesquisas com alunos vêm se proliferando. Assim como pesquisas com pais e observações dos professores, em que outros educadores mais experientes observam um professor durante uma aula.

"Distritos escolares perceberam que não deveriam focar exclusivamente notas de testes", diz Thomas Kane, professor de educação em Harvard. Um estudo de três anos realizado pela Fundação Bill & Melinda Gates demonstra o valor tanto do Big quanto do Small Data para a educação. Os autores analisaram se modelos baseados em notas, pesquisas com alunos ou observação de professores foram melhores para medir quais professores mais aprimoraram o aprendizado dos alunos. Quando reuniram estas três medidas em uma única pontuação composta, obtiveram os melhores resultados. "Cada medida acrescenta algo de valor", concluiu o relatório.

Na verdade, foi exatamente quando descobri que muitas operações de Big Data usam Small Data para preencher as lacunas que fui a Ocala, Flórida, para me encontrar com Jeff Seder. Ele é o guru dos cavalos, formado em Harvard, que utilizou as lições aprendidas a partir de um enorme conjunto de dados para prever o sucesso de *American Pharoah*.

Depois de compartilhar todos os arquivos digitais e a matemática comigo, Seder admitiu que tinha outra arma: Patty Murray.

Murray, assim como Seder, é muito inteligente e tem as melhores credenciais — um diploma de Bryn Mawr. Ela também trocou a cidade de Nova York pela vida no campo. "Gosto mais de cavalos do que de pessoas", admite Murray. Mas Murray é um pouco mais tradicional em suas abordagens para avaliação de cavalos. Ela, como muitos agentes de cavalos, examina pessoalmente os animais, observando como andam, checando cicatrizes e lesões e interrogando os proprietários.

Murray então colabora com Seder na escolha final dos cavalos que querem recomendar. Ela tem faro para detectar problemas nos cavalos, do tipo que os dados de Seder, apesar de serem o conjunto de dados mais inovador e importante já coletado sobre cavalos, ainda ignoram.

Estou prevendo uma revolução baseada nas revelações do Big Data. Mas isto não significa que podemos simplesmente lançar mão dos dados em qualquer pergunta. E o Big Data não elimina a necessidade de todas as outras maneiras que os seres humanos desenvolveram ao longo do milênio para compreender o mundo. Eles se complementam.

8

MAIS DADOS, MAIS PROBLEMAS? O QUE NÃO DEVEMOS FAZER

Às vezes, o poder do Big Data é tão impressionante que é assustador. Ele levanta questões éticas.

O PERIGO DAS CORPORAÇÕES EMPODERADAS

Recentemente, três economistas — Oded Netzer e Alain Lemaire, ambos da Universidade de Columbia, e Michal Herzenstein, da Universidade de Delaware — procuravam por maneiras de prever a probabilidade de pagamento de empréstimo por um tomador. Os acadêmicos utilizaram dados do Prosper, um site de empréstimos direto entre usuários. Potenciais tomadores escrevem uma breve descrição do motivo pelo qual precisam do dinheiro e por que têm uma boa chance de pagar a dívida, e os potenciais credores decidem se vão ou não emprestar o dinheiro. De modo geral, cerca de 13% dos tomadores se tornam inadimplentes.

Ocorre que a linguagem usada pelos potenciais tomadores de empréstimo é um forte preditor de sua probabilidade de pagar a dívida. E ela é um importante indicativo mesmo utilizando o controle de outras informações relevantes que os credores foram capazes de obter sobre os potenciais devedores, incluindo avaliação de crédito e renda.

As dez frases que os pesquisadores descobriram que são comumente usadas quando se pede o empréstimo estão listadas abaixo. Cinco delas se correlacionam positivamente com o pagamento da dívida. Cinco, negativamente. Em outras palavras, cinco tendem a ser usadas por pessoas em que você pode confiar, cinco, por quem não se pode confiar. Veja se adivinha qual é qual.

Deus
menor taxa de juros
deduzido o imposto
prometo
pagarei
hospital
livre de dívida
formado
pagamento mínimo
obrigado

Você pode pensar — ou pelo menos esperar — que uma pessoa educada, abertamente religiosa, que promete honra com sua palavra estaria entre as com maior probabilidade de pagar o empréstimo. Mas na verdade não é o

caso. Este tipo de pessoa, os dados mostram, tem menos probabilidade do que a média de pagar o débito.

Veja agora as frases agrupadas pela probabilidade de pagar o empréstimo.

TERMOS USADOS EM PEDIDOS DE EMPRÉSTIMO POR PESSOAS COM MAIOR PROBABILIDADE DE HONRAR O PAGAMENTO

livre de dívida deduzido o imposto formado

menor taxa de juros pagamento mínimo

TERMOS USADOS EM PEDIDOS DE EMPRÉSTIMO POR PESSOAS COM MAIOR PROBABILIDADE DE NÃO PAGAR

Deus pagarei hospital

prometo obrigado

Antes de discutirmos as implicações éticas deste estudo, vamos analisar, com a ajuda dos autores, o que isto revela sobre as pessoas. O que podemos inferir das palavras nas diferentes categorias?

Primeiro, consideremos a linguagem que sugere que alguém tem maior probabilidade de pagar o empréstimo. Frases como "menor taxa de juros" ou "deduzido o imposto" indicam um certo nível de sofisticação financeira da parte do tomador, então não é de surpreender que se correlacionem com alguém com maior probabilidade de honrar a dívida. Além disso, se ele ou ela fala de realizações positivas, como ser "formado" em uma universidade e estar "livre de dívidas", também tem maior probabilidade de pagar o empréstimo.

Agora analisemos a linguagem que indica a improbabilidade de alguém pagar a dívida. Geralmente, se alguém diz que vai lhe pagar, é porque não vai. Quanto mais confiante a promessa, maior a probabilidade de a quebrar. Alguém que escreve: "Prometo que pagarei, juro por Deus", está entre aqueles com menor probabilidade de pagar a dívida. Apelar para sua compaixão — explicando que precisa do dinheiro por que um parente está no "hospital" — também significa menor probabilidade de pagar. Na verdade, mencionar um familiar — marido, mulher, filho, mãe ou pai — é um sinal de que alguém não vai honrar o pagamento. Outra palavra que indica inadimplência é "explicar", significando que se a pessoa está tentando explicar por que conseguirá pagar o empréstimo, provavelmente não o fará.

Os autores não têm uma teoria de por que demonstrar gratidão é uma evidência de inadimplência.

Resumindo, segundo esses pesquisadores, fornecer um plano detalhado de como serão capazes de fazer os pagamentos e mencionar outras obrigações cumpridas no passado são evidências de inadimplência. Fazer promessas e apelar para a compaixão é um sinal claro de que alguém não vai lhe pagar. Independentemente dos motivos — ou o que isso nos diz sobre a natureza humana, que fazer promessas é um sinal claro de que alguém, na verdade, não vai fazer coisa alguma —, os acadêmicos consideraram o teste uma informação extremamente valiosa na predição da inadimplência. Alguém que menciona Deus tinha 2,2 vezes mais probabilidade de não pagar. Isto estava entre os indicadores únicos mais sólidos de que alguém vai dar o calote.

Mas os autores também acreditam que o estudo levanta questões éticas. Apesar de ser apenas um estudo acadêmico, algumas empresas relatam utilizar os dados online para aprovar empréstimos. Isto é aceitável? Queremos viver em um mundo em que as empresas usam as palavras que escrevemos na internet para prever se vamos ou não pagar um empréstimo? É, no mínimo, repugnante — e, muito provavelmente, assustador.

Um consumidor buscando um empréstimo em um futuro próximo pode precisar se preocupar não só com seu histórico financeiro, mas também com sua atividade online. E pode ser julgado por fatores que parecem absurdos — como, por exemplo, se usa expressões como "obrigado" ou invoca a "Deus". E quanto a uma pessoa que realmente

precisa ajudar sua irmã no hospital e que certamente pagará o empréstimo? Parece terrível puni-la porque, em média, pessoas que alegam precisar do dinheiro para despesas médicas frequentemente mentem. Um mundo que funciona desta maneira começa a parecer terrivelmente distópico.

Esta é uma questão ética: as empresas têm o direito de julgar nossa adequação para seus serviços com base em critérios preditivos abstratos, mas estatísticos, não relacionados a seus serviços?

Deixando de lado o mundo das finanças, analisemos algumas implicações mais graves, como, por exemplo, as práticas de contratação. Empregadores cada vez mais investigam a mídia social ao avaliar candidatos a empregos. Isto pode não levantar questões éticas se eles estiverem apenas procurando por evidências de calúnia ou revelação de segredos de empregadores prévios. Pode até haver alguma justificativa para recusar a contratação de alguém cujos posts no Facebook ou no Instagram sugiram uso excessivo de álcool. Mas e se o empregador encontrar um indicador aparentemente inofensivo que se correlacione com algo importante para ele?

Pesquisadores da Universidade de Cambridge e da Microsoft submeteram 58 mil usuários do Facebook dos Estados Unidos a uma variedade de testes sobre sua personalidade e inteligência. Eles descobriram que as curtidas no Facebook frequentemente se correlacionam com QI, extroversão e consciência. Por exemplo, pessoas que curtem Mozart, tempestades de raios e batatas fritas no Facebook tendem a ter QIs maiores. Pessoas que curtem motos Harley-Davidson, o grupo de música country Lady Antebellum, ou a página “Eu Adoro Ser Mãe” tendem a ter QIs menores. Algumas destas correlações são produto da maldição da dimensionalidade. Se testarmos um universo grande o suficiente de coisas, algumas irão se correlacionar aleatoriamente. Mas alguns interesses apresentam uma correlação legítima com o QI.

Contudo, seria injusto se uma pessoa inteligente que por acaso curte Harleys não conseguisse um emprego condizente com suas habilidades porque foi, sem perceber, marcada como tendo baixa inteligência.

Para ser justo, este não é um problema totalmente novo. As pessoas há muito tempo são julgadas por fatores não diretamente relacionados ao desempenho no emprego — a firmeza do aperto de mão, o capricho de suas roupas. Mas um perigo da revolução dos dados é que, à medida que mais aspectos de nossa vida são quantificados, esses julgamentos indiretos se tornam cada vez mais herméticos e ainda mais invasivos. Previsões melhores levam a uma discriminação mais sútil e nefasta.

Dados melhores também levam a outra forma de discriminação, a que economistas chamam de discriminação de preços. Empresas tentam o tempo todo descobrir qual preço devem cobrar por mercadorias ou serviços. Idealmente, querem cobrar o máximo que os clientes estão dispostos a pagar. Desta forma, obtêm o maior lucro possível.

A maioria dos negócios normalmente acaba escolhendo um preço que todo mundo pague. Mas, por vezes, sabem que membros de determinado grupo, em média, pagariam mais. É por isso que cinemas cobram mais dos clientes de meia-idade — que estão no ápice de sua capacidade financeira — do que de estudantes ou idosos, e é por isso que as companhias aéreas normalmente cobram mais de passageiros de última hora. Eles discriminam por preço.

O Big Data permite que as empresas aperfeiçoem substancialmente seu conhecimento sobre o quanto seus clientes estão dispostos a pagar — e assim lucram mais com certos grupos de pessoas. O Optimal Decisions Group foi pioneiro na utilização de ciência de dados para prever quanto os consumidores se dispõem a pagar por um seguro. Como? Usando a metodologia que discutimos anteriormente neste livro. Eles encontraram clientes anteriores semelhantes aos que atualmente procuravam por seguros — e verificaram qual a recompensa mais alta que estavam dispostos a aceitar. Em outras palavras, conduziram uma busca por um dúplice. Este tipo de busca é divertido se ajuda a prever se um jogador de beisebol voltará à antiga grandiosidade. Ela é ótima se ajuda a curar a doença de alguém. Mas e se a busca pelo dúplice auxilia uma empresa a arrancar até seu último centavo? Isto não é nada divertido. Meu irmão perdulário teria todo o direito de reclamar se cobrassem mais dele por um serviço do que de mim, que sou pão-duro.

Jogos de azar são outra área em que a habilidade de focar determinados clientes é potencialmente perigosa. Os grandes cassinos usam algo parecido com a busca por dúplices para entender melhor seu público. O objetivo? Obter o maior lucro possível — para ter certeza de que mais de seu dinheiro vá para seus cofres.

Funciona da seguinte forma. Cada jogador, creem os cassinos, tem um “ponto sensível”. É o montante de perdas que fará com que o cliente se afaste do cassino por um longo período. Suponha, por exemplo, que o “ponto sensível” de Helen seja US$3 mil. Isto significa que se ela perder US$3 mil, o cassino perde a cliente, talvez por semanas ou meses. Se Helen perder US$2.999, não vai ficar feliz. Afinal, quem gosta de perder dinheiro? Mas não ficará tão arrasada a ponto de não voltar na noite seguinte.

Imagine por um momento que você é o gerente do cassino. E suponha que Helen apareça para jogar nas máquinas caça-níqueis. Qual é o resultado ideal? Obviamente, você quer que Helen chegue o mais perto possível de

seu “ponto sensível” sem o alcançar. Quer que ela perca US$2.999, o suficiente para obter um lucro alto, mas não o bastante para ela não voltar amanhã para jogar.

Como fazer isto? Bem, existem maneiras de fazer com que Helen pare de jogar depois de perder certa quantia. Você pode oferecer refeições grátis, por exemplo. Faça uma oferta sedutora o bastante e ela sairá das máquinas e irá ao restaurante.

Mas existe um grande desafio nesta abordagem. Como saber qual é o “ponto sensível” de Helen? O problema é que as pessoas têm diferentes “pontos sensíveis”. Para Helen, é US$3 mil. Para John, pode ser US$2 mil. Para Ben, US$26 mil. Se você convencer Helen a parar de jogar depois de perder US$2 mil, deixa de arrecadar todo o lucro possível. Se esperar demais — depois de ela ter perdido US$3 mil —, perdeu a cliente por um tempo. Além do mais, Helen pode não querer lhe dizer qual seu “ponto sensível”. Pode ser que nem ela mesma saiba qual é.

Então, o que fazer? Se chegou até aqui neste livro, provavelmente deve ter adivinhado a resposta. Use a ciência de dados. Aprenda tudo que puder sobre um determinado número de clientes — idade, gênero, CEP e comportamento de jogo. E, a partir do comportamento de jogo — seus prêmios, perdas, idas e vindas —, estime seu “ponto sensível”.

Reúna toda a informação que souber sobre Helen e encontre jogadores parecidos com ela — seus dúplices, mais ou menos. Depois, descubra o quanto de perdas podem suportar. Provavelmente, o valor será o mesmo de Helen. De fato, é isto o que o cassino de Harrah faz, utilizando uma empresa de armazenamento de Big Data, a Terabyte, para o ajudar.

Scott Gnau, gerente-geral da Terabyte, explica, no excelente livro *Super Crunchers*, de Ian Ayres, o que os gerentes do cassino fazem quando veem um cliente habitual chegar perto de seu “ponto sensível”: “Eles se aproximam e dizem ‘Vejo que você está tendo um dia difícil, sei que gosta de nosso restaurante. Gostaria que levasse sua mulher para jantar por nossa conta agora mesmo.’”

Isto pode parecer um ato de extrema generosidade: um jantar grátis. Mas, na verdade, eles atendem aos próprios interesses. O cassino tenta fazer com que os clientes desistam antes que percam tanto que desapareçam por um longo período. Em outras palavras, os administradores usam análises de dados sofisticadas para obter o máximo de dinheiro possível de seus clientes, em longo prazo.

Temos razão em temer que uma utilização cada vez maior dos dados online, melhor proporcionará aos cassinos, companhias de seguro, financeiras e outras entidades empresariais poder demais sobre nós.

Por outro lado, o Big Data também permite que consumidores revidem contra as empresas que fazem cobranças indevidas ou que entregam produtos de má qualidade.

Uma importante arma são os sites, como o Yelp, que publica críticas de restaurantes e outros serviços. Um recente estudo conduzido pelo economista Michael Luca, de Harvard, demonstrou em que extensão as empresas acabam à mercê dos críticos do Yelp. Comparando as críticas aos dados de vendas no estado de Washington, ele descobriu que uma estrela a menos no Yelp resulta em uma queda de 5% a 9% no faturamento do restaurante.

Os consumidores também são auxiliados em suas lutas contra as empresas pelos sites de comparação de compras — como o Kayak e o Booking.com. Como discutido em *Freakonomics*, quando um site na internet começou a reportar os preços que as diferentes empresas cobravam por seguros de vida, eles despencaram drasticamente. Se uma seguradora cobrasse caro demais, os clientes saberiam e procurariam outra empresa. O total da economia para os consumidores? Um bilhão de dólares por ano.

Os dados da internet, em outras palavras, dizem às empresas quais clientes evitar e quais explorar. Eles também mostram aos clientes de quais empresas fugir e quais estão tentando extorqui-los. O Big Data até hoje já ajudou os dois lados da disputa entre consumidores e empresas. Temos que garantir que esta batalha continue justa.

O PERIGO DOS GOVERNOS EMPODERADOS

Quando seu ex-namorado apareceu em sua festa de aniversário, Adriana Donato sabia que ele estava perturbado. Ela sabia que estava nervoso. Sabia que andara depressivo. Quando ele a convidou para dar uma volta de carro, havia uma coisa que Adriana, uma estudante de zoologia de 20 anos, não sabia. Ela não sabia que seu ex-namorado de 22 anos de idade, James Stoneham, passara as três semanas anteriores pesquisando informações sobre como matar alguém e as leis aplicáveis ao homicídio, intercaladas com buscas ocasionais sobre Adriana.

Se ela soubesse disto, tudo indica que não entraria no carro. Provavelmente, não teria sido esfaqueada até a morte naquela noite.

No filme *Minority Report — A Nova Lei*, videntes colaboram com a polícia para impedir crimes antes que aconteçam. Será que o Big Data deveria ser disponibilizado aos departamentos de polícia para impedir a ocorrência

de crimes? Será que Adriana Donato pelo menos deveria ter sido avisada sobre as buscas sinistras do ex-namorado? A polícia deveria ter interrogado James Stoneham?

Primeiro, é preciso reconhecer que há evidências crescentes de que as buscas no Google relacionadas a crimes se correlacionam de fato com ações criminosas. Christine Ma-Kellams, Flora Or, Ji Hyun Baek e Ichiro Kawachi demonstraram que as buscas no Google ligadas a suicídio têm forte correlação com suas taxas em um estado. Além disto, Evan Soltas e eu mostramos que as buscas islamofóbicas semanais — como “eu odeio muçulmanos” ou “matar muçulmanos” — se correlacionaram com crimes de ódio naquela semana. Se mais pessoas realizam buscas dizendo que desejam fazer algo, é porque mais pessoas realmente colocam em prática.

Então, o que deveríamos fazer com essas informações? Uma ideia simples e bastante incontroversa: podemos utilizar os dados de uma determinada área para alocar recursos. Se uma cidade tem um enorme número de buscas relacionadas ao suicídio, é possível aumentar a atenção para o suicídio naquele local. O governo municipal ou organizações sem fins lucrativos poderiam exibir campanhas publicitárias explicando para as pessoas onde buscar ajuda, por exemplo. Da mesma forma, se uma cidade experimenta um enorme aumento nas buscas por “matar muçulmanos”, os departamentos de polícia poderiam alterar o esquema de patrulha das ruas; enviando mais policiais para proteger uma mesquita, por exemplo.

Mas uma atitude deveríamos relutar em tomar: perseguir pessoas antes que um crime seja cometido. Isto parece, em primeiro lugar, invasão de privacidade. Há uma enorme diferença ética entre um governo levantar os dados de busca de milhares ou centenas de milhares de pessoas e um departamento de polícia que obtém os dados de um indivíduo. Há um salto ético gigantesco entre proteger a mesquita local e revistar a casa de alguém. Há uma grande distância entre fazer campanhas por causa do suicídio e internar alguém em um hospital psiquiátrico contra sua vontade.

A razão para sermos extremamente cautelosos ao usar dados de indivíduos, porém, vai além até da ética. Há uma razão nos próprios dados, também. Há um gigantesco passo para a ciência de dados entre tentar prever as ações de uma cidade e as de um indivíduo.

Voltemos a questão do suicídio, por um momento. Todos os meses, ocorrem cerca de 3,5 milhões de buscas no Google nos Estados Unidos relacionadas ao suicídio, sendo a maioria delas indicativas de ideias suicidas — buscas como “suicida”, “cometer suicídio” e “como se suicidar”. Em outras palavras, todo mês, há mais do que uma busca relacionada ao suicídio para cada cem norte-americanos. Isto me faz recordar da citação do filósofo Friedrich Nietzsche: “A ideia de suicídio é um grande consolo: com ela superamos muitas noites más.” Os dados de busca no Google mostram o quanto isto é verdadeiro, o quanto a ideia de suicídio é comum. Entretanto, todos os meses, ocorrem menos de quatro mil suicídios nos Estados Unidos. A ideia suicida é incrivelmente comum. O suicídio, não. Assim, não faria muito sentido que policiais passassem a bater à porta de todo mundo que já divulgou na internet intenções de explodir os próprios miolos — pelo simples motivo de que a polícia não teria tempo de fazer mais nada.

Ou pense naquelas buscas islamofóbicas incrivelmente cruéis. Em 2015, houve aproximadamente 12 mil buscas nos Estados Unidos por “matar muçulmanos”. Ocorreram 12 homicídios de muçulmanos reportados como crimes de ódio. Claramente, a vasta maioria das pessoas que fazem esta busca horripilante não a colocou em prática.

Um pouco de matemática é capaz de explicar a diferença entre prever o comportamento de um indivíduo e de uma cidade. Vejamos um simples experimento mental. Suponha que existam um milhão de pessoas em uma cidade e uma mesquita. Se alguém não faz uma busca por “matar muçulmanos”, há apenas 1 chance em 1 milhão de que realmente atacará uma mesquita. Suponha que alguém faça uma busca por “matar muçulmanos”, esta chance aumenta drasticamente, para 1 em 10 mil. Imagine que a islamofobia tenha explodido no país e as buscas por “matar muçulmanos” tenham aumentado de 100 para 1.000.

Nesta situação, a matemática mostra que as chances de uma mesquita ser atacada aumentam cerca de 5 vezes, de 2% para 10%. Mas a probabilidade de um indivíduo que pesquisou a frase “matar muçulmanos” realmente atacar uma mesquita permanece de apenas 1 em 10 mil.

A resposta adequada nesta situação não é mandar para a cadeia todas as pessoas que buscaram por “matar muçulmanos”. Nem é visitar suas casas. Existe uma pequena chance de que qualquer uma dessas pessoas em particular cometa um crime. A resposta apropriada, porém, seria proteger aquela mesquita, que agora tem 10% de probabilidade de ser atacada.

Obviamente, muitas buscas apavorantes não levam a ações terríveis.

Contudo, é ao menos teoricamente possível que alguns tipos de buscas sugiram uma probabilidade razoavelmente alta de uma consequência nefasta. Teoricamente, pelo menos, é possível que cientistas de dados possam no futuro criar um modelo que consiga descobrir que as buscas de James Stoneham relacionadas a Adriana Donato eram causa de preocupação real.

Em 2014, ocorreram cerca de 6 mil buscas para a frase exata “como matar sua namorada” e 400 assassinatos de namoradas. Se todos os assassinos fizeram esta busca exata com antecedência, isto significaria que 1 a cada 15 pessoas que buscaram “como matar sua namorada” colocaram a ideia em prática. Obviamente, muitas, provavelmente a maioria, das pessoas que assassinaram suas namoradas não pesquisaram exatamente esta frase. Isto indica que a verdadeira probabilidade desta busca resultar em homicídio é menor, possivelmente muito menor.

Mas se os cientistas de dados pudessem criar um modelo que mostrasse que as ameaças contra um indivíduo específico eram de, digamos, 1 em 100, poderíamos querer fazer algo com esta informação. Ao menos, a pessoa sob ameaça deveria ter o direito de ser informada que existe uma chance de 1 em 100 de que seja assassinada por uma determinada pessoa.

Em geral, entretanto, temos que ser muito cautelosos ao utilizar os dados de busca para prever crimes a nível individual. Os dados nos dizem claramente que ocorrem inúmeras buscas desta natureza, mas que raramente resultam em ações funestas. E não houve, até agora, provas de que o governo prevê uma ação danosa específica, com alta probabilidade de se concretizar, apenas examinando buscas. Assim, temos que ser realmente cuidadosos em permitir que o governo intervenha a nível individual com base em dados de busca. Não apenas por motivos éticos ou legais, mas também, pelo menos por ora, por razões atinentes à ciência de dados.

CONCLUSÃO

QUANTAS PESSOAS TERMINAM OS LIVROS QUE LEEM?

Depois de assinar o contrato de publicação deste livro, tinha uma visão clara de como o livro deveria ser estruturado. No início, você deve lembrar, descrevi uma cena de um evento familiar. Meus familiares debateram sobre minha sanidade e tentaram descobrir por que, aos 33 anos, eu não conseguia encontrar a mulher certa.

A conclusão deste livro, então, praticamente se escreveria sozinha. Encontraria a mulher ideal e me casaria. Melhor ainda, usaria Big Data para a conhecer. Talvez pudesse analisar dados de processos judiciais. Então a história toda chegaria ao ápice na conclusão, que descreveria meu casamento e serviria de carta de amor para minha nova esposa.

Infelizmente, a vida não correspondeu à minha expectativa. Ficar trancado em meu apartamento e evitar o mundo enquanto escrevia provavelmente não ajudou em minha vida amorosa. E eu, infelizmente, ainda preciso encontrar uma esposa. E o mais importante, precisava de uma nova conclusão.

Vasculhei muitos de meus livros favoritos tentando encontrar o que é preciso em uma boa conclusão. As melhores conclusões, constatei, são as que trazem à tona um ponto importante que estava o tempo todo ali, pairando logo abaixo da superfície. Para este livro, este ponto é: a ciência social está começando a se tornar uma ciência real. E esta nova e real ciência está pronta para melhorar nossas vidas.

No início da Parte II, discuti a crítica de Karl Popper a Sigmund Freud. Popper, observei, não pensava que a visão excêntrica de Freud do mundo era científica. Mas não mencionei algo sobre a crítica de Popper. Ela foi muito além de um simples ataque a Freud. Popper não considerava que *qualquer* cientista social fosse especialmente científico. Popper simplesmente não ficou impressionado com o rigor daquilo que esses “pretensos” cientistas faziam.

O que motivou a cruzada de Popper? Quando interagia com os mais iminentes intelectuais de sua época — os melhores físicos, historiadores e psicólogos — Popper observou uma impressionante diferença. Quando os físicos conversavam, Popper acreditava no que faziam. Claro, às vezes eles cometiam erros. Sim, às vezes se deixavam levar por tendências subconscientes. Mas os físicos trabalhavam em um processo que obviamente descobria verdades profundas sobre o mundo, que culminou na Teoria da Relatividade, de Einstein. Quando os mais famosos cientistas sociais do mundo falavam, por outro lado, Popper achava que estava ouvindo um monte de jargão afetado e sem sentido.

Popper não foi nem de longe a única pessoa a fazer esta distinção. Quase todo mundo concorda que físicos, biólogos e químicos são cientistas de verdade. Eles empregam rigorosos experimentos para descobrir como o mundo físico funciona. Em contrapartida, muitas pessoas acham que economistas, sociólogos e psicólogos são cientistas indulgentes que andam por aí despejando jargões superficiais para evoluir na carreira.

Considerando que algum dia isto tenha sido verdade, a revolução do Big Data mudou tudo. Se Karl Popper estivesse vivo hoje e assistisse a uma apresentação de Raj Chetty, Jesse Shapiro, Esther Duflo ou (só para me deixar feliz) minha, tenho fortes suspeitas de que ele não teria a mesma reação de antes. Para ser franco, ele estaria mais propenso a questionar se os defensores da teoria das cordas de hoje são cientistas de verdade ou se estão apenas praticando ginástica mental hedonista.

Se um filme violento chega aos cinemas, a criminalidade aumenta ou diminui? Se mais pessoas são expostas a um anúncio, surgem mais usuários do produto? Se um time de basquete vence quando um rapaz tem 20 anos, é mais

provável que torça pelo time aos 40? Todas estas são perguntas diretas, com respostas objetivas do tipo sim ou não. E em montanhas de dados genuínos, podemos encontrá-las.

Isto é ciência, não pseudociência.

O que não significa que a revolução da ciência social ocorrerá na forma de leis simples e estáticas.

Marvin Minsky, falecido cientista do MIT e um dos primeiros a estudar a possibilidade de inteligência artificial, sugeriu que a psicologia perdeu o rumo ao tentar copiar a física, que obtinha sucesso encontrando leis elementares que se mantivessem verdadeiras em todos os momentos e em todos os lugares.

O cérebro humano, Minsky sugeria, não pode se sujeitar a essas leis. O cérebro é um sistema elaborado de soluções — uma parte corrigindo os erros em outras. O sistema econômico e político pode ser similarmente complexo.

Por esta razão, é improvável que a revolução da ciência social venha sob fórmulas elegantes, como $E = MC^2$. Na verdade, se alguém alega uma revolução da ciência social com base em fórmulas claras, é melhor se manter cético.

A revolução ocorrerá gradualmente, estudo por estudo, descoberta por descoberta. Lentamente, teremos um melhor entendimento dos sistemas complexos da mente humana e da sociedade.

Uma conclusão apropriada deve resumir a questão, mas também apontar o caminho para as coisas que virão.

Para este livro, isto é fácil. Os conjuntos de dados discutidos aqui são revolucionários, mas ainda há muito a ser explorado. Há muito mais para ser aprendido. Sinceramente, a esmagadora maioria dos acadêmicos tem ignorado a explosão de dados provocada pela era digital. Os pesquisadores mais famosos do mundo no campo sexual continuam utilizando métodos provados e comprovados. Eles perguntam a algumas centenas de sujeitos sobre seus desejos; não pedem dados de sites como o PornHub. Os mais conceituados linguistas do mundo analisam textos individuais; basicamente ignoram os padrões revelados em bilhões de livros. As metodologias ensinadas a estudantes de graduação em psicologia, ciência política e sociologia continuam, em grande parte, intocadas pela revolução digital. O território vasto e quase inexplorado aberto pela explosão dos dados tem sido relegado a um pequeno número de professores inovadores, estudantes rebeldes e curiosos.

Isto irá mudar.

Para cada conceito que discuti neste livro, há centenas de ideias igualmente importantes prontas para ser exploradas. A pesquisa discutida aqui é a ponta da ponta do iceberg, um arranhão no arranhão da superfície.

Então, o que mais vem por aí?

Primeiro, uma expansão radical da metodologia que era usada em um dos estudos mais bem-sucedidos em saúde pública de todos os tempos. Em meados do século XIX, John Snow, físico britânico, estava interessado em descobrir o que provocou um surto de cólera em Londres.

Sua ideia engenhosa: ele mapeou todos os casos de cólera na cidade. Ao concluir, descobriu que a doença estava basicamente agrupada ao redor de uma determinada bomba-d'água. Isto sugeria que se disseminava através da água contaminada, refutando a ideia vulgarmente aceita de que se espalhava pelo ar de má qualidade.

O Big Data — e a possibilidade de individualização que oferece — facilita este tipo de estudo. Para qualquer doença, podemos explorar os dados de busca no Google ou outros dados digitais de saúde. É possível descobrir se há bolsões no mundo em que a prevalência desta doença é incomumente alta ou baixa. Então, podemos ver o que esses lugares têm em comum. Será algo no ar? Na água? Nas regras sociais?

É possível fazer a mesma análise para enxaquecas ou pedras nos rins. Podemos rastrear ansiedade e depressão, Mal de Alzheimer, câncer pancreático, pressão alta, dores nas costas, constipação e sangramentos nasais. Podemos fazer esta pesquisa para tudo. A análise que Snow fez uma vez, somos capazes de fazer quatrocentas vezes (algo em que já comecei a trabalhar).

Isto — pegar um método simples e utilizar Big Data para realizar uma análise centenas de vezes em um curto período de tempo — é ciência em grande escala. Sim, as ciências social e comportamental certamente serão escalonadas. Focar as condições de saúde ajudará essas ciências a ganhar escala. Outra coisa que ajudará o processo: o teste A/B. Discutimos o teste A/B no contexto dos negócios para fazer com que usuários cliquem em títulos ou anúncios — e este tem sido o uso predominante da metodologia. Mas ele pode ser usado para revelar coisas essenciais — e socialmente valiosas —, para além de setas que fazem com que as pessoas cliquem em um anúncio.

Benjamin F. Jones é um economista da Universidade de Northwestern que usa o teste A/B para ajudar crianças a aprender melhor. Ele é um dos criadores da plataforma EDU STAR, que permite que as escolas testem aleatoriamente diferentes planos de aulas.

Existem muitas empresas no campo de softwares de educação. Com a EDU STAR, os alunos se conectam a um computador e são aleatoriamente expostos a diferentes planos de aulas. Depois, fazem pequenos testes para ver o quanto aprenderam o conteúdo. Em outras palavras, as escolas descobrem qual software funciona melhor para ajudar os alunos a compreender a matéria.

Como todas as grandes plataformas de testes A/B, a EDU STAR já produziu resultados surpreendentes. Um dos planos de aulas que conquistou a preferência de muitos educadores incluía um software que utilizava jogos para ajudar a ensinar frações aos alunos. Certamente, se transformarmos matemática em um jogo, os alunos se divertirão mais, aprenderão mais e se sairão melhor em provas. Certo? Errado. Os alunos que tiveram aulas de frações por meio de um jogo tiveram notas piores do que aqueles que aprenderam de uma forma mais tradicional.

Fazer com que crianças aprendam mais é um uso empolgante e socialmente benéfico do teste desenvolvido no Vale do Silício para fazer com que as pessoas cliquem em mais anúncios. Assim como sua utilização para que as pessoas durmam mais.

O norte-americano médio dorme 6,7 horas toda noite. A maioria gostaria de dormir mais. Mas às 23h, o canal de esportes está a todo vapor ou o YouTube faz seu apelo irresistível. Então, o sono tem que esperar. Jawbone, uma empresa de dispositivos usáveis com centenas de milhares de clientes, conduz milhares de testes para tentar descobrir intervenções que ajudem seus usuários a fazer o que querem: dormir mais cedo.

Jawbone obteve gigantesco sucesso com um objetivo duplo. Primeiro, pediu aos clientes para se comprometerem com um objetivo "não muito ambicioso". Enviou mensagens como esta: "Parece que você não tem dormido muito nos últimos 3 dias. Por que não tenta ir dormir às 23h30 esta noite? Sabemos que você normalmente se levanta às 8h." Então, os usuários têm a opção de clicar em "Aceito".

Segundo, às 22h30, Jawbone envia outra mensagem: "Decidimos que você dormiria às 23h30. São 22h30. Por que não começa agora?"

Jawbone descobriu que esta estratégia resultou em vinte e três minutos a mais de sono. Eles não foram capazes de fazer os usuários irem para a cama às 22h30, mas conseguiram que fossem dormir mais cedo.

Obviamente, cada uma das partes desta estratégia tinha que ser otimizada através de diversos experimentos. Comece o objetivo original cedo demais — peça aos usuários que se comprometam a ir para a cama às 23h —, e poucos concordarão. Peça para irem dormir à meia-noite e o ganho será pequeno.

Jawbone usou o teste A/B para descobrir o equivalente do sono à seta apontando para a direita da Google. Mas em vez de obter alguns cliques extras para os parceiros publicitários da Google, o resultado foram minutos a mais de descanso para norte-americanos exaustos.

Na verdade, o campo inteiro da psicologia pode utilizar as ferramentas do Vale do Silício para melhorar drasticamente suas pesquisas. Espero ansiosamente o primeiro artigo de psicologia que, em vez de detalhar poucos experimentos feitos com alguns estudantes, mostre os resultados de mil testes A/B rápidos.

Os dias dos acadêmicos dedicando meses para recrutar um pequeno número de estudantes para realizar um único teste chegarão ao fim. Em vez disso, acadêmicos utilizarão dados digitais para testar algumas centenas ou alguns milhares de ideias em poucos segundos. Seremos capazes de aprender muito mais em um tempo muito menor.

Textos usados como dados nos revelarão muito mais. Como as ideias são disseminadas? Como novas palavras são criadas? Como as palavras desaparecem? Como surgem as piadas? Por que algumas palavras são engraçadas e outras não? Como surgem os dialetos? Aposto que, dentro de vinte anos, teremos descobertas profundas sobre todas estas questões.

Penso que devemos considerar a utilização do comportamento dos jovens online — devidamente protegido pelo anonimato — como um complemento aos testes tradicionais para verificar como estão aprendendo e se desenvolvendo. Como escrevem? Demonstram sinais de dislexia? Têm interesses intelectuais maduros? Eles têm amigos? Existem pistas para todas estas perguntas em milhares de toques de teclados que cada um deles realiza todos os dias.

E existe outra área, não trivial, de onde surgem muitas outras novas percepções.

Na música *Shattered* ["Destruído", em tradução livre], dos Rolling Stones, Mick Jagger descreve tudo que torna a cidade de Nova York, a famosa Big Apple, tão mágica. Risos. Alegria. Solidão. Ratos. Percevejos. Orgulho. Ganância. Pessoas vestidas em sacos de papel. Mas Jagger dedica a maior parte das palavras para o que a torna realmente especial: "Sexo e sexo e sexo e sexo".

Na Big Apple, assim como no Big Data. Graças à revolução digital, novas descobertas surgem na área da saúde. Sono. Aprendizado. Psicologia. Linguagem. Além de sexo e sexo e sexo e sexo.

Uma pergunta que atualmente pesquiso: quantas dimensões de sexualidade existem? Normalmente pensamos nas pessoas como homo ou heterossexuais. Mas a sexualidade é nitidamente muito mais complexa. Entre homo e heterossexuais, as pessoas têm tipos — alguns homens gostam de "louras", outros, de "morenas", por exemplo. Estas preferências são tão fortes quanto o gênero? Outra pergunta que investigo: de onde vêm as preferências

sexuais? Assim como é possível descobrir os principais anos que determinam o time de beisebol para que torcem ou as visões políticas, podemos descobrir agora os anos determinantes para as preferências sexuais. Para descobrir essas respostas, você terá que comprar meu próximo livro, temporariamente intitulado *Everybody (Still) Lies* [“Todo Mundo (Ainda) Mente”, em tradução livre].

A existência da pornografia — e os dados que a acompanham — é um desenvolvimento revolucionário na ciência da sexualidade humana.

Levou tempo para que as ciências naturais começassem a mudar nossas vidas — para descobrir a penicilina, criar os satélites e computadores. Pode demorar um pouco antes que o Big Data conduza as ciências sociais e comportamentais a importantes avanços na forma como amamos, aprendemos ou vivemos. Mas acredito que estes avanços se aproximam. Espero que você consiga ver pelo menos os contornos de seu desenvolvimento neste livro. Espero, na verdade, que alguns que estão lendo este livro ajudem a criá-los.

Para escrever uma conclusão apropriada, um autor deveria pensar por que afinal escreveu o livro. Qual é o objetivo que pretende alcançar?

Acredito que a principal razão para escrever este livro tenha sido o resultado de uma das experiências mais impressionantes de minha vida. Um pouco mais de uma década atrás, o livro *Freakonomics* foi lançado. O surpreendente best-seller descreveu a pesquisa de Steven Levitt, um economista premiado da Universidade de Chicago mencionado muitas vezes neste livro. Levitt foi um “economista selvagem”, que parecia ser capaz de usar os dados para responder a qualquer pergunta que sua mente excêntrica pudesse imaginar: lutadores de sumô são desleais? Competidores de programas de televisão são preconceituosos? Os corretores de imóveis conseguem para você os mesmos tipos de negócios que conseguem para si mesmos?

Eu acabara de sair da faculdade, com formação em filosofia, com pouca noção do que pretendia fazer da vida. Depois de ler *Freakonomics*, eu sabia. Queria fazer o mesmo que Steven Levitt. Pretendia mergulhar em montanhas de dados para descobrir como o mundo *realmente* funcionava. Eu seguiria seus passos, decidi, e faria meu doutorado em economia.

Tanta coisa mudou no intervalo de vinte anos. Alguns estudos de Levitt continham erros de codificação. Levitt fez algumas declarações politicamente incorretas sobre o aquecimento global. *Freakonomics* perdeu o prestígio nos círculos intelectuais.

Mas acredito que, excluídos alguns equívocos, os anos foram muito favoráveis para um grande ponto que Levitt pretendia demonstrar. Levitt nos mostrou que uma combinação de curiosidade, criatividade e dados poderia aprimorar drasticamente nossa compreensão do mundo. Havia histórias escondidas em dados prontas para ser contadas, e isto se mostrou verdadeiro de novo e de novo.

Espero que este livro tenha o mesmo efeito nas pessoas que *Freakonomics* teve em mim. Espero que haja jovens lendo este livro agora que estejam um tanto confusos sobre o que desejam para suas vidas. Se você tem um pouco de habilidade estatística, muita criatividade e curiosidade, entre para o mundo da análise de dados.

Este livro, na verdade, se me permite tamanha ousadia, pode ser encarado como um *Freakonomics* avançado. Uma das principais diferenças entre os estudos discutidos no *Freakonomics* e neste livro é a ambição. Na década de 1990, quando Levitt ganhou fama, não havia muitos dados disponíveis. Levitt se satisfez em buscar perguntas excêntricas, onde havia dados disponíveis. Ele basicamente ignorou grandes perguntas onde não existiam dados. Hoje, porém, com tantos dados disponíveis sobre quase todos os assuntos, faz sentido buscar perguntas grandes e profundas que extraiam a essência do que significa ser humano.

O futuro da análise de dados é brilhante. O próximo Kinsey, tenho forte convicção, será um cientista de dados. O próximo Foucault o será. O próximo Freud, também. O próximo Marx será um cientista de dados. O próximo Salk pode muito bem também sê-lo.

De qualquer forma, esta foi minha tentativa de fazer algumas das coisas que uma conclusão adequada faz. Mas excelentes conclusões, vim a perceber, fazem muito mais. Muito, muito mais. Uma excelente conclusão precisa ser irônica. Ela precisa ser tocante. Uma grande conclusão precisa ser sagaz e divertida. Precisa ser profunda, engraçada e triste. Uma excelente conclusão precisa, em uma ou duas sentenças, mostrar uma ideia que resuma tudo que veio antes, tudo que vem depois. Precisa fazer isto com um ápice dramático, único — uma guinada. Um grande livro precisa terminar com uma exclamação provocativa, inteligente e divertida.

Agora talvez seja um bom momento para falar um pouco sobre o processo de escrever. Não sou um escritor especialmente eloquente. Este livro tem apenas cerca de 75 mil palavras, o que é relativamente pouco para um

tópico rico como este.

Mas o que me falta em extensão, compenso em determinação. Passei cinco meses, e fiz 47 rascunhos, escrevendo minha primeira coluna sobre sexo no *New York Times*, que continha 2 mil palavras. Alguns capítulos neste livro precisaram de 60 rascunhos. Sou capaz de passar horas tentando descobrir a palavra certa para uma frase em uma nota de rodapé.

Vivi como eremita a maior parte do último ano. Apenas eu e meu computador. Morei na parte mais badalada de Nova York e quase nunca saía. Esta é, na minha opinião, minha obra-prima, a melhor ideia que terei em minha vida. E estava disposto a sacrificar qualquer coisa para fazê-la bem. Queria ser capaz de defender cada palavra neste livro. Meu telefone está repleto de e-mails que esqueci de responder, convites virtuais que nunca abri, mensagens do Bumble que ignorei.*

Depois de trinta meses de trabalho árduo, finalmente pude mandar um rascunho quase completo. Uma parte, porém, estava faltando: a conclusão.

Expliquei à minha editora, Denise, que a conclusão poderia demorar mais alguns meses. Disse que meu palpite era de seis meses. A conclusão, na minha opinião, é a parte mais importante do livro. E eu estava apenas aprendendo o que torna uma conclusão excelente. Desnecessário dizer que Denise não ficou satisfeita.

Então, um dia, um amigo me enviou um e-mail com um estudo de Jordan Ellenberg. Ellenberg, um matemático da Universidade de Wisconsin, estava curioso sobre quantas pessoas realmente terminam os livros que leem. Ele pensou em uma maneira engenhosa de testar isto usando Big Data. A Amazon relata quantas pessoas citam diversas linhas finais de livros. Ellenberg percebeu que poderia comparar com que frequência as citações eram extraídas do início e do final de um livro. Isso poderia dar uma noção geral da propensão de o leitor chegar ao fim. Pela sua pesquisa, mais de 90% dos leitores terminaram a leitura de *O Pintassilgo (The Goldfinch),* de Donna Tartt. Em contrapartida, apenas 7% terminaram de ler a obra-prima do economista ganhador do Prêmio Nobel, Daniel Kahneman, *Rápido e Devagar — Duas Formas de Pensar (Thinking, Fast e Slow)*. Menos de 3%, esta metodologia estimou, leu até o fim o livro, muito comentado e elogiado do economista Thomas Piketty, O *Capital no Século XXI (Capital in the 21st Century)*. Em outras palavras, as pessoas tendem a não terminar obras de economistas.

Um dos pontos deste livro é que seguimos o Big Data *para onde quer* que nos leve e agimos de acordo. Espero que a maioria dos leitores absorva cada palavra e tente detectar padrões conectando as páginas finais ao que aconteceu antes. Mas, não importa o quanto me esmerei para aprimorar minha prosa, muitas pessoas lerão apenas as primeiras cinquenta páginas e seguirão com suas vidas.

Assim, concluo este livro da única forma apropriada: seguindo os dados, o que as pessoas realmente fazem, não o que dizem. Vou beber uma cerveja com meus amigos e parar de trabalhar nesta maldita conclusão. Poucos de vocês, o Big Data me diz, ainda estão aqui.

* Como todo mundo mente, você deve estar questionando a veracidade desta história. Talvez eu não seja de fato um trabalhador obsessivo. Talvez eu não tenha trabalhado com tanto afinco neste livro. Talvez, como muitas pessoas, eu tenha exagerado o quanto trabalho. Talvez meus trinta meses de "trabalho árduo" incluíram meses inteiros em que não trabalhei. Talvez eu não tenha vivido como um eremita. Talvez, se checar meu perfil no Facebook, veria fotos minhas com meus amigos no período de suposto isolamento. Ou talvez eu tenha sido um eremita, mas isto não tenha sido imposição própria. Talvez tenha passado muitas noites sozinho, incapaz de trabalhar, esperando em vão que alguém entrasse em contato. Talvez ninguém tenha me enviado mensagens no Bumble. Todo mundo mente. Nenhum contador de histórias é digno de confiança.

AGRADECIMENTOS

Este livro foi um trabalho de equipe.

Estas ideias foram desenvolvidas enquanto eu era estudante em Harvard, cientista de dados na Google e escritor do *New York Times*.

Hal Varian, com quem trabalhei na Google, foi uma grande influência nas ideias deste livro. O que posso dizer, Hal está constantemente vinte anos à frente de seu tempo. Seu livro *A Economia da Informação,* escrito com Carl Shapiro, basicamente prevê o futuro. E seu artigo *Predicting the Present* [“Prevendo o Presente”, em tradução livre] escrito com Hyunyoung Choi foi um dos grandes responsáveis pela revolução do Big Data nas ciências sociais que descrevo neste livro. Ele é também um mentor gentil e maravilhoso, como muitas das pessoas que já trabalharam com ele podem confirmar. Um comportamento típico de Hal é executar a maior parte do trabalho em um artigo que você está escrevendo com ele e depois insistir que seu nome apareça como primeiro autor. A combinação de genialidade e generosidade de Hal é algo que raramente encontrei antes.

Minha escrita e ideias evoluíram sob a orientação de Aaron Retica, que tem sido meu editor em cada artigo escrito para a coluna do *New York Times*. Aaron é um erudito. Ele de alguma forma consegue saber tudo sobre música, história, esportes, política, sociologia, economia e só Deus sabe o que mais. Ele é responsável por grande parte de tudo de bom escrito na coluna do *Times* que leva meu nome. Outros membros da equipe desta coluna incluem Bill Marsh, cujos gráficos continuam a me encantar, Kevin McCarthy e Gita Daneshjoo. Este livro inclui trechos desta coluna, reproduzidos com autorização.

Steven Pinker, que gentilmente concordou em escrever o prefácio, há muito tempo é um herói para mim. Ele estabeleceu os parâmetros para um livro moderno sobre ciência social — um explorador engajado dos elementos básicos da natureza humana, revelando as melhores pesquisas em uma gama de disciplinas. Este é o parâmetro que busco alcançar por toda minha vida.

Minha tese, a partir da qual este livro foi desenvolvido, foi escrita sob a paciente e brilhante orientação de Alberto Alesina, David Cutler, Ed Glaeser e Lawrence Katz.

Denise Oswald é uma editora fantástica. Se quiser saber o quanto é maravilhosa, basta comparar este livro com meu primeiro rascunho — na verdade, você não poderá fazer isso pois nunca mostrarei aquele manuscrito horroroso para mais ninguém. Agradeço ainda o restante da equipe da HarperCollins, incluindo Michael Barrs, Lynn Grady, Lauren Janiec, Shelby Meizlik e Amber Oliver.

Eric Lupfer, meu agente, que viu o potencial do meu projeto desde o início, foi essencial na criação da proposta e ajudou durante todo o processo.

Pela verificação de fatos sensacional, agradeço a Melvis Acosta.

Outras pessoas com quem aprendi muito em minha vida acadêmica e profissional incluem Susan Athey, Shlomo Benartzi, Jason Bordoff, Danielle Bowers, David Broockman, Bo Cowgill, Steven Delpome, John Donohue, Bill Gale, Claudia Goldin, Suzanne Greenberg, Shane Greenstein, Steve Grove, Mike Hoyt, David Laibson, A.J. Magnuson, Dana Maloney, Jeffrey Oldham, Peter Orszag, David Reiley, Jonathan Rosenberg, Michael Schwarz, Steve Scott, Rich Shavelson, Michael D. Smith, Lawrence Summers, Jon Vaver, Michael Wiggins e Qing Wu.

Agradeço a Tim Requarth e a NeuWrite por me ajudarem a desenvolver minha escrita.

Por colaborarem na interpretação dos estudos, agradeço a Christopher Chabris, Raj Chetty, Matt Gentzkow, Solomon Messing e Jesse Shapiro.

Pedi a Emma Pierson e a Katia Sobolski se poderiam me aconselhar em um capítulo deste livro. Por alguma razão fora de minha compreensão, elas se ofereceram para ler o livro todo — e me deram sábios conselhos em cada parágrafo.

Minha mãe, Esther Davidowitz, leu o livro inteiro em diversas ocasiões e ajudou imensamente em seu aprimoramento. Ela também me ensinou, por exemplo, que eu deveria seguir minha curiosidade, não importasse para onde me levasse. Quando estava sendo entrevistado para um cargo acadêmico, um professor me perguntou: “O que sua mãe acha do trabalho que você faz?” A ideia era que minha mãe pudesse se envergonhar pela minha pesquisa sobre sexo e outros assuntos tabu. Mas eu sempre soube que ela se orgulhava de mim por eu ter seguido minha curiosidade, para onde quer que ela tivesse me levado.

Muitas pessoas leram partes do livro e fizeram comentários muito proveitosos. Agradeço a Eduardo Acevedo, Coren Apicella, Sam Asher, David Cutler, Stephen Dubner, Christopher Glazek, Jessica Goldberg, Lauren Goldman, Amanda Gordon, Jacob Leshno, Alex Peysakhovich, Noah Popp, Ramon Roullard, Greg Sobolski, Evan Soltas, Noah Stephens-Davidowitz, Lauren StephensDavidowitz e Jean Yang. Na verdade, Jean foi minha melhor amiga enquanto escrevia o livro, e por isto também lhe sou imensamente grato.

Por ajudar na coleta de dados, agradeço a Brett Goldenberg, James Rogers e Mike Williams, da MindGeek, e Rob McQuown e Sam Miller, da Baseball Prospectus.

Sou profundamente grato ao apoio financeiro da Alfred Sloan Foundation.

Em determinado ponto, ao escrever este livro, fiquei empacado, perdido e quase abandonei o projeto. Então fiz uma viagem ao campo com meu pai, Mitchell Stephens. Ao longo de uma semana, ele me colocou de volta nos trilhos. Ele me levou para caminhadas nas quais discutimos sobre amor, morte, sucesso, felicidade e escrever — e então sentou-se comigo repassando cada frase do livro. Eu não teria concluído o livro sem ele.

Todos os erros remanescentes são, obviamente, meus.

REFERÊNCIAS

INTRODUÇÃO: OS CONTORNOS DE UMA REVOLUÇÃO

A grande maioria dos eleitores norte-americanos não se importava que Barack Obama. Katie Fretland, “Gallup: Race Not Important to Voters”, The Swamp. *Chicago Tribune*: junho de 2008.

Análise de Berkeley. Alexandre Mas e Enrico Moretti, “Racial Bias in the 2008 Presidential Election“, *American Economic Review* 99, n° 2 (2009).

sociedade pós-racial. Em 12 de novembro de 2009, em um episódio do seu programa, Lou Dobbs falou em uma “sociedade pós-partidária, pós-racial”. Em 27 de janeiro de 2010, em um episódio de seu programa, Chris Matthews falou que o Presidente Obama era “aparentemente pós-racial”. Para outros exemplos, veja Michael C. Dawson e Lawrence D. Bobo, “One Year Later e the Myth of a Post-Racial Society”, *Du Bois Review: Social Science Research on Race* 6, n° 2 (2009).

Analisei os dados da General Social Survey. Detalhes sobre todos estes cálculos são encontrados em meu site, sethsd.com [conteúdo em inglês], no csv chamado “Sex Data”. Dados da General Social Survey estão em http://gss.norc.org/ [conteúdo em inglês].

menos de 600 milhões de camisinhas. Dados fornecidos para o autor.

buscas e assinaturas para o Stormfront. Análise do autor dos dados do Google Trends. Também coletei dados de todos os membros do Stormfront, como discutido no artigo de Seth Stephens-Davidowitz, “The Data of Hate”, *New York Times*, 13 de julho de 2014, SR4. Os dados atinentes podem ser baixados em sethsd.com [conteúdo em inglês] na seção de dados rotulada “Stormfront”.

mais buscas por “nigger president” [presidente crioulo] do que por “first black president” [primeiro presidente negro]. Análise do autor do Google Trends. Os estados em que isto ocorreu incluem Kentucky, Louisiana, Arizona e Carolina do Norte.

rejeitado por cinco jornais acadêmicos. O artigo foi posteriormente publicado por Seth Stephens-Davidowitz, “The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data”, *Journal of Public Economics* 118 (2014). Mais detalhes sobre a pesquisa podem ser encontrados lá. Os dados estão também em meu site, sethsd.com [conteúdo em inglês], na seção de dados chamada “Racism”.

único fator que se correlacionou adequadamente. “O mais forte correlato que encontrei para o apoio a Trump são as buscas no Google pela ‘palavra com n’. Outros pesquisadores reportaram o mesmo” (tuíte em 28 de fevereiro de 2016). Veja também o artigo de Nate Cohn, “Donald Trump’s Strongest Supporters: A New Kind of Democrat”, *New York Times*, 31 de dezembro de 2015, A3.

as que faziam mais buscas no Google pelo termo “nigger”. Observe que, como a medição é feita em porcentual de buscas no Google, não é majoritariamente superior em lugares com grandes populações ou que fazem muitas buscas. Observe, ainda, que algumas diferenças neste mapa e no mapa de apoio a Trump têm explicações óbvias. Trump perdeu popularidade no Texas e no Arkansas, pois estes são os estados natais de seus adversários, Ted Cruz e Mike Huckabee.

Estes são dados de pesquisa da Civis Analytics de dezembro de 2015. Dados reais de votação são menos úteis aqui, pois são altamente influenciados pelo momento de realização da prévia e pelo formato da votação. Os mapas foram reimpressos com autorização do *New York Times*.

2,5 milhões de terabytes de dados. “Bringing Big Data to the Enterprise”, IBM, https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html [conteúdo em inglês].

agulhas, se preferir, em palheiros cada vez maiores. Nassim M. Taleb, “Beware the Big Errors of ‘Big Data’”, *Wired*, 8 de fevereiro de 2013, http://www.wired.com/2013/02/big-data-means-big-errors-people [conteúdo em inglês].

nem as buscas racistas nem as assinaturas no Stormfront. Examinei como o racismo na internet mudou em partes do país com alta e baixa exposição à Grande Recessão. Analisei as taxas de busca no Google pela palavra “nigger(s)” e as assinaturas no Stormfront. Os dados pertinentes podem ser baixados em sethsd.com [conteúdo em inglês], nas seções de dados denominadas “Racial Animus” e “Stormfront”.

Mas as buscas no Google que refletem ansiedade. Seth Stephens-Davidowitz, “Fifty States of Anxiety”, *New York Times*, 7 de agosto de 2016, SR2. Observe que, enquanto as buscas no Google nos fornecem amostras muito maiores, este padrão é consistente com as provas das pesquisas. Veja, por exemplo, William C. Reeves et al., “Mental Illness Surveillance Among Adults in the United States”, *Morbidity e Mortality Weekly Report Supplement* 60, n° 3 (2011).

buscas por piadas. Isto é discutido no artigo de Seth Stephens-Davidowitz, “Why Are You Laughing?”, *New York Times*, 15 de maio de 2016, SR9. Os dados pertinentes podem ser baixados em sethsd.com [conteúdo em inglês], na seção de dados chamada “Jokes.”

“meu marido quer que eu o amamente”. Isto é discutido no artigo de Seth Stephens-Davidowitz, “What Do Pregnant Women Want?”, New York Times, 17 de maio de 2014, SR6.

buscas pornográficas por imagens de mulheres amamentando homens. Análise do autor nos dados de PornHub.

Mulheres fazem quase o mesmo número. Isto é discutido no artigo de Seth Stephens-Davidowitz, "Searching for Sex", New York Times, 25 de janeiro de 2015, SR1.

"poemas para minha esposa grávida". Stephens-Davidowitz, "What Do Pregnant Women Want?"

disse Friedman. Entrevistei Jerry Friedman pelo telefone em 27 de outubro de 2015.

amostra de todos seus dados. Hal R. Varian, "Big Data: New Tricks for Econometrics", *Journal of Economic Perspectives* 28, n° 2 (2014).

CAPÍTULO 1: SUA INTUIÇÃO FALHA

A melhor ciência de dados, na verdade, é surpreendentemente intuitiva. Falo sobre o campo da análise de dados que conheço — ciência de dados que tenta explicar e prever o comportamento humano. Não falo de inteligência artificial que tenta, digamos, dirigir um carro. Estas metodologias, embora utilizem as ferramentas descobertas pelo cérebro humano, são menos fáceis de entender.

quais sintomas previam o câncer de pâncreas. John Paparrizos, Ryan W. White e Eric Horvitz, "Screening for Pancreatic Adenocarcinoma Using Signals from Web Search Logs: Feasibility Study e Results", *Journal of Oncology Practice* (2016).

O clima de inverno superou todas. Esta pesquisa é discutida no artigo de Seth Stephens-Davidowitz, "Dr. Google Will See You Now", *New York Times*, 11 de agosto de 2013, SR12.

o maior conjunto de dados jamais coletado sobre relacionamento humano. Lars Backstrom e Jon Kleinberg, "Romantic Partnerships e the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook" em *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2014).

as pessoas reiteradamente classificam. Daniel Kahneman, "Rápido e Devagar: Duas Formas de Pensar" (New York: Farrar, Straus e Giroux, 2011).

a asma provoca aproximadamente 70 vezes mais mortes. Entre 1979 e 2010, em média, 55,81 norte-americanos morreram em tornados e 4216,53, de asma. Veja Annual U.S. Killer Tornado Statistics, National Weather Service, http://www.spc.noaa.gov/climo/torn/fatalmap.php e Trends in Asthma Morbidity e Mortality, American Lung Association, Epidemiology e Statistics Unit [conteúdo em inglês].

Patrick Ewing. Meus vídeos favoritos de Ewing são "Patrick Ewing's Top 10 Career Plays", YouTube, postado em 18 de setembro de 2015, https://www.youtube.com/watch?v=Y29gMuYymv8 e "Patrick Ewing Knicks Tribute" YouTube, postados em 12 de maio de 2006, https://www.youtube.com/watch?v=8T2l5Emzu-I [conteúdo em inglês].

"… basquete como questão de vida e morte". S. L. Price, "Whatever Happened to the White Athlete?", *Sports Illustrated*, 8 de dezembro de 1997.

pesquisa na internet. Esta foi a Pesquisa de Consumidor Google que realizei em 22 de outubro de 2013. Perguntei: "Onde você acha que a maioria dos jogadores da NBA nasceu?" As duas opções eram "bairros pobres" e "bairros de classe média"; 59,7% dos respondentes escolheram "bairros pobres".

o primeiro nome de uma pessoa negra é um indicador de seu contexto socioeconômico. Roland G. Fryer Jr. e Steven D. Levitt, "The Causes e Consequences of Distinctively Black Names". Quarterly Journal of Economics 119, n° 3 (2004).

afro-americanos nascidos na década de 1980. Centers for Disease Control and Prevention, "Health, United States, 2009" Table 9, Nonmarital Childbearing, by Detailed Race and Hispanic Origin of Mother, and Maternal Age: Estados Unidos, Anos 1970 – 2006.

Chris Bosh… Chris Paul. "Not Just a Typical Jock: Miami Heat Forward Chris Bosh's Interests Go Well Beyond Basketball", Palm BeachPost.com, 15 de fevereiro de 2011, http://www.palmbeachpost.com/news/sports/basketball/not-just-a-typical-jock-miami-heat-forward-chris-b/nLp7Z/ [conteúdo em inglês]; Dave Walker, "Chris Paul's Family to Compete on 'Family Feud,' nola.com, 31 de outubro de 2011, http://www.nola.com/tv/index.ssf/2011/10/chris_pauls_family_to_compete.html [conteúdo em inglês].

dez centímetros mais alto. "Why Are We Getting Taller as a Species?", Scientific American, http://www.scientificamerican.com/article/why-are-we-getting-taller/ [conteúdo em inglês]. Curiosamente, os norte-americanos pararam de ficar mais altos. Amanda Onion, "Why Have Americans Stopped Growing Taller?", ABC News, 3 de julho de 2016, http://abcnews.go.com/Technology/story?id=98438&page=1 [conteúdo em inglês]. Defendo que uma das razões pelas quais houve um aumento expressivo de jogadores da NBA estrangeiros é que outros países estão alcançando a média de altura dos Estados Unidos. O número de norte-americanos com mais de 2,13m na NBA aumentou dezesseis vezes entre 1946 e 1980, enquanto os norte-americanos cresciam. Desde então esta taxa se estabilizou, conforme os norte-americanos foram parando de crescer. Enquanto isso, o número de atletas com mais de 2,13m de outros países aumentou substancialmente. O maior aumento nos jogadores internacionais, descobri, é de homens extremamente altos de países como Turquia, Espanha e Grécia, em que houve notáveis aprimoramentos na saúde infantil e na altura de adultos nos últimos anos.

norte-americanos de comunidades subdesenvolvidas. Carmen R. Isasi et al., "Association of Childhood Economic Hardship with Adult Height e Adult Adiposity among Hispanics/Latinos: The HCHS/SOL SocioCultural Ancillary Study", *PloS One* 11, n° 2 (2016); Jane E. Miller e Sanders Korenman, "Poverty e Children's Nutritional Status in the United States, *American Journal of Epidemiology* 140, n° 3 (1994); Harry J. Holzer, Diane Whitmore Schanzenbach, Greg J. Duncan e Jens Ludwig, "The Economic Costs of Childhood Poverty in the United States", *Journal of Children e Poverty* 14, n° 1 (2008).

a altura média de um norte-americano é 1,79m. Cheryl D. Fryar, Qiuping Gu e Cynthia L. Ogden, "Anthropometric Reference Data for Children e Adults: United States, 2007–2010", *Vital e Health Statistics Series* 11, n°. 252 (2012).

um a cada cinco chegue à NBA. Pablo S. Torre, "Larger Than Real Life" ["Maior que a Vida Real", em tradução livre] *Sports Illustrated*, 4 de julho de 2011.

classe média, com ambos os pais. Tim Kautz, James J. Heckman, Ron Diris, Bas Ter Weel e Lex Borghans, "Fostering e Measuring Skills: Improving Cognitive e Non-Cognitive Skills to Promote Lifetime Success", National Bureau of Economic Research Working Paper 20749, 2014.

Wrenn pulava mais alto. Desmond Conner, "For Wrenn, Sky's the Limit" *Hartford Courant*, 21 de outubro de 1999.

Mas Wrenn. A história de Doug Wrenn é contada por Percy Allen, "Former Washington e O'Dea Star Doug Wrenn Finds Tough Times", *Seattle Times*, 29 de março de 2009.

"… Doug Wrenn está morto…". Ibid.

Jordan poderia ter sido um garoto difícil. Melissa Isaacson, "Portrait of a Legend" ESPN.com, 9 de setembro de 2009, http://www.espn.com/chicago/columns/story?id=4457017&columnist=isaacson_melissa. Uma boa biografia de Jordan é o livro de Roland Lazenby, *Michael Jordan: The Life*, Boston: Back Bay Books, 2015 [conteúdo em inglês].

Seu pai era. Barry Jacobs, "High-Flying Michael Jordan Has North Carolina Cruising Toward Another NCAA Title" *People*, 19 de março de 1984.

a vida de Jordan está repleta de histórias de sua família o desviando das armadilhas. Isaacson, "Portrait of a Legend."

admissão no Hall da Fama do Basquete. Discurso de Michael Jordan na consagração do Hall da Fama do Basquete, YouTube, postado em 21 de fevereiro de 2012, https://www.youtube.com/watch?v=XLzBMGXfK4c [conteúdo em inglês]. O aspecto mais interessante do discurso de Jordan não é o fato de ele ser tão efusivo sobre seus pais, e sim ainda sentir a necessidade de apontar o desprezo sofrido no início da carreira. Talvez uma vida inteira de obsessão com o desprezo seja necessária para se tornar o maior jogador de basquete de todos os tempos.

LeBron James foi entrevistado. "I'm LeBron James from Akron, Ohio", vídeo do YouTube, postado em 20 de junho de 2013, https://www.youtube.com/watch?v=XceMbPVAggk [conteúdo em inglês].

CAPÍTULO 2: FREUD ESTAVA CERTO?

alimento em formato fálico. Codifiquei alimentos como sendo fálicos quando eram significativamente mais longos do que largos e geralmente arredondados. Considerei pepino, milho, cenoura, berinjela, abóbora e banana. Os dados e o código podem ser encontrados em sethsd.com [conteúdo em inglês].

erros de digitação coletados por pesquisadores da Microsoft. O conjunto de dados pode ser baixado em https://www.microsoft.com/en-us/download/details.aspx?id=52418 [conteúdo em inglês]. Os pesquisadores pediram aos usuários do Amazon Mechanical Turk para descrever imagens. Eles analisaram os registros de teclas digitadas e anotaram todas as vezes que alguém corrigiu uma palavra. Mais detalhes no artigo de Yukino Baba e Hisami Suzuki, "How Are Spelling Errors Generated e Corrected? A Study of Corrected e Uncorrected Spelling Errors Using Keystroke Logs", Proceedings of the Fiftieth Annual Meeting of the Association for Computational Linguistics, 2012. Os dados, códigos e uma descrição mais detalhada desta pesquisa são encontrados em sethsd.com [conteúdo em inglês].

Considere todas as buscas na forma "Eu quero fazer sexo com meu/minha…". Os dados completos — aviso: descritivo — são os seguintes:

"EU QUERO FAZER SEXO COM…"

	BUSCAS NO GOOGLE MENSAIS COM A FRASE EXATA
minha mãe	720
meu filho	590
minha irmã	590
meu primo	480
meu pai	480
meu namorado	480
meu irmão	320
minha filha	260
meu/minha amigo/a	170
minha namorada	140

desenhos animados pornôs. Por exemplo, "pornô" é uma das palavras mais comuns incluídas nas buscas no Google em vários programas de animação extremamente populares, conforme mostrado abaixo.

DESENHOS E PORNOGRAFIA
(BUSCAS NO GOOGLE MAIS COMUNS PARA DIVERSOS DESENHOS ANIMADOS)

family guy pornô	assista os simpsons	**futurama pornô**	scooby doo jogos
family guy episódios	**os simpsons pornô**	futurama leela	scooby doo filmes
family guy grátis	os simpsons online	futurama episódios	**scooby doo pornô**
assista family guy	os simpsons filme	futurama online	scooby doo velma

babás. Com base nos cálculos do autor, essas são as ocupações mais populares das mulheres nas buscas pornográficas de homens, separadas pela faixa etária dos homens:

OCUPAÇÕES DE MULHERES EM BUSCAS PORNOGRÁFICAS FEITAS POR HOMENS, POR FAIXA ETÁRIA DOS HOMENS

	18–24	25–64	65+
1.	Babá	Babá	Babá
2.	Professora	Instrutora de Yoga	Líder de torcida
3.	Instrutora de Yoga	Professora	Médica
4.	Líder de torcida	Líder de torcida	Professora
5.	Médica	Corretora de imóveis	Corretora de imóveis
6.	Prostituta	Médica	Enfermeira
7.	Corretora de imóveis	Prostituta	Instrutora de Yoga
8.	Enfermeira	Secretária	Secretária
9.	Secretária	Enfermeira	Prostituta

CAPÍTULO 3: OS DADOS REINVENTADOS

algoritmos operando. Matthew Leising, "HFT Treasury Trading Hurts Market When News Is Released", Bloomberg Markets, 16 de dezembro de 2014; Nathaniel Popper, "The Robots Are Coming for Wall Street", *New York Times Magazine*, 28 de fevereiro de 2016, MM56; Richard Finger, "High Frequency Trading: Is It a Dark Force Against Ordinary Human Traders e Investors?" *Forbes*, 30 de setembro de 2013, http://www.forbes.com/sites/richardfinger/2013/09/30/high-frequency-trading-is-it-a-dark-force-against-ordinary-human-traders-and-investors/#50875fc751a6 [conteúdo em inglês].

Alan Krueger. Entrevistei Alan Krueger por telefone em 8 de maio de 2015.

importantes indicadores da velocidade de disseminação da gripe. O artigo original foi de Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski e Larry Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data", Nature 457, n° 7232 (2009). As falhas no modelo original foram discutidas no artigo de David Lazer, Ryan Kennedy, Gary King e Alessandro Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis" Science 343, n° 6176 (2014). Um modelo corrigido é apresentado no artigo de Shihao Yang, Mauricio Santillana e S. C. Kou, "Accurate Estimation of Influenza Epidemics Using Google Search Data Via ARGO", *Proceedings of the National Academy of Sciences* 112, n° 47 (2015).

quais buscas se relacionam mais intimamente aos preços de moradia. Seth Stephens-Davidowitz e Hal Varian, "A Hands-on Guide to Google Data", mimeo, 2015. Veja também Marcelle Chauvet, Stuart Gabriel e Chandler Lutz, "Mortgage Default Risk: New Evidence from Internet Search Queries", *Journal of Urban Economics* 96 (2016).

Bill Clinton. Sergey Brin e Larry Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine" Seventh International World Wide Web Conference, 14–18 de abril de 1998, Brisbane, Austrália.

sites pornográficos. John Battelle, A pesquisa: "How Google e Its Rivals Rewrote the Rules of Business e Transformed Our Culture" (Nova York: Penguin, 2005).

fornecendo uma opinião. Uma boa discussão sobre isto pode ser encontrada no artigo de Steven Levy, *In the Plex: How Google Thinks, Works, e Shapes Our Lives* (New York: Simon & Schuster, 2011).

"Venda sua casa". Esta citação também foi incluída no artigo Joe Drape, "Ahmed Zayat's Journey: Bankruptcy e Big Bets" *New York Times*, 5 de Junho de 2015, A1. Entretanto, o artigo atribui incorretamente a citação a Seder. Ela foi feita, na verdade, por outro membro de sua equipe.

Conheci Seder. Entrevistei Jeff Seder e Patty Murray em Ocala, Florida, de 12 de junho de 2015 até 14 de junho de 2015.

Aproximadamente um terço. A razão do fracasso dos cavalos de corrida são as estimativas aproximadas de Jeff Seder, com base em anos de experiência no setor.

centenas de cavalos morrem. Supplemental Tables of Equine Injury Database Statistics for Thoroughbreds, http://jockeyclub.com/pdfs/eid_7_year_tables.pdf [conteúdo em inglês].

grande parte por fraturas nas pernas. "Postmortem Examination Program", California Animal Health e Food Laboratory System, 2013.

Ainda assim, mais de três quartos não vencem uma grande corrida. Avalyn Hunter, "A Case for Full Siblings" *Bloodhorse*, 18 de abril de 2014, http://www.bloodhorse.com/horse-racing/articles/115014/a-case-for-full-siblings [conteúdo em inglês].

Earvin Johnson III. Melody Chiu, "E. J. Johnson Loses 50 Lbs. Since Undergoing Gastric Sleeve Surgery", *People*, 1° de outubro de 2014.

LeBron James, cuja mãe tem 1,65m. Eli Saslow, "Lost Stories of LeBron, Part 1" ESPN.com, 17 de outubro de 2013, http://www.espn.com/nba/story/_/id/9825052/how-lebron-james-life-changed-fourth-grade-espn-magazine [conteúdo em inglês].

The Green Monkey. Veja artigo de Sherry Ross, "16 Million Dollar Baby", *New York Daily News*, 12 de março de 2006; e de Jay Privman, "The Green Monkey, Who Sold for $16M, Retired", ESPN.com, 12 de fevereiro de 2008, http://www.espn.com/sports/horse/news/stor?id=3242341 [conteúdo em inglês]. Um vídeo do leilão está disponível no YouTube "$16 Million Horse", postado em 1° de novembro de 2008, https://www.youtube.com/watch?v=EyggMC85Zsg [conteúdo em inglês].

Uma fraqueza da tentativa do Google de prever a gripe. Sharad Goel, Jake M. Hofman, Sébastien Lahaie, David M. Pennock e Duncan J. Watts, "Predicting Consumer Behavior with Web Search", Proceedings of the National Academy of Sciences 107, n° 41 (2010).

Pop-Tarts de morango. Constance L. Hays, "What Wal-Mart Knows About Customers' Habits", *New York Times*, 14 de novembro de 2004.

"Funcionou muito bem.". Entrevistei Orley Ashenfelter por telefone em 27 de outubro de 2016.

Eles estudaram centenas de encontros rápidos entre heterossexuais. Daniel A. McFarland, Dan Jurafsky e Craig Rawlings, "Making the Connection: Social Bonding in Courtship Situations", *American Journal of Sociology* 118, n° 6 (2013).

Leonard Cohen certa vez deu a seu sobrinho o seguinte conselho para conquistar uma mulher. Jonathan Greenberg, "What I Learned From My Wise Uncle Leonard Cohen", *Huffington Post*, 11 de novembro de 2016.

as palavras usadas por centenas de milhares de posts no Facebook. H. Andrew Schwartz et al., "Personality, Gender, e Age in the Language of Social Media: The Open-Vocabulary Approach", *PloS One 8*, n° 9 (2013). O artigo também desmembra as formas que as pessoas falam com base em como pontuam em testes de personalidade. Veja o que eles encontraram:

textos contidos em milhares de livros e roteiros de filmes. Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth e Peter Sheridan Dodds, “The Emotional Arcs of Stories Are Dominated by Six Basic Shapes”, *EPJ Data Science* 5, n° 1 (2016).

quais tipos de histórias são compartilhadas. Jonah Berger e Katherine L. Milkman, “What Makes Online Content Viral?”, *Journal of Marketing Research* 49, n° 2 (2012).

por que algumas publicações tendem para a esquerda. Esta pesquisa é toda mostrada no artigo de Matthew Gentzkow e Jesse M. Shapiro, “What Drives Media Slant? Evidence from U.S. Daily Newspapers” *Econometrica 78*, n° 1 (2010). Embora fossem apenas estudantes de doutorado quando este projeto começou, Gentzkow e Shapiro são agora estrelas da economia. Gentzkow, agora professor em Stanford, venceu a 2014 John Bates Clark Medal, outorgada ao melhor economista abaixo de 40 anos. Shapiro, agora professor em Brown, é editor do prestigiado *Journal of Political Economy*. O artigo conjunto deles sobre mídia tendenciosa está entre os mais citados de ambos.

Rupert Murdoch’s. A propriedade de Murdoch do conservador *New York Post* poderia ser explicada pelo fato de Nova York ser tão grande, que é capaz de apoiar jornais de múltiplos pontos de vista. Entretanto, parece muito claro que o *Post* vem perdendo dinheiro reiteradamente. Veja, por exemplo, de Joe Pompeo, “How Much Does the ‘New York Post’ Actually Lose?” *Politico*, 30 de Agosto de 2013, http://www.politico.com/media/story/2013/08/how-much-does-the-new-york-post-actually-lose-001176 [conteúdo em inglês].

me disse Shapiro. Entrevistei Matt Gentzkow e Jesse Shapiro em 16 de agosto de 2015, no Royal Sonesta Boston.

muitas escolas de ensino médio escanearam seus anuários. Kate Rakelly, Sarah Sachs, Brian Yin e Alexei A. Efros, “A Century of Portraits: A Visual Historical Record of American High School Yearbooks”, artigo apresentado na International Conference on Computer Vision, 2015. As fotos são reimpressas com permissão dos outros autores.

pessoas fotografadas copiavam as pessoas nos quadros. Veja, por exemplo, Christina Kotchemidova, “Why We Say ‘Cheese’: Producing the Smile in Snapshot Photography”, *Critical Studies in Media Communication* 22, n° 1 (2005).

medir o PIB com base na quantidade de luzes acesas à noite nesses países. J. Vernon Henderson, Adam Storeygard e David N. Weil, “Measuring Economic Growth from Outer Space”, American Economic Review 102, n° 2 (2012).

A estimativa do PIB foi 90% maior. Kathleen Caulderwood, “Nigerian GDP Jumps 89% as Economists Add in Telecoms, Nollywood” *IBTimes*, 7 de abril de 2014, http://www.ibtimes.com/nigerian-gdp-jumps-89-economists-add-telecoms-nollywood-1568219 [conteúdo em inglês].

disse Reisinger. Entrevistei Joe Reisinger pelo telefone em 10 de junho de 2015.

US$50 milhões. Leena Rao, “SpaceX e Tesla Backer Just Invested $50 Million in This Startup”, *Fortune*, 24 de setembro de 2015.

CAPÍTULO 4: SORO DIGITAL DA VERDADE

Um importante artigo de 1950. Hugh J. Parry e Helen M. Crossley, “Validity of Responses to Survey Questions” *Public Opinion Quarterly* 14, 1 (1950).

Uma pesquisa recente questionou os graduandos da Universidade de Maryland. Frauke Kreuter, Stanley Presser e Roger Tourangeau, “Social Desirability Bias in CATI, IVR, and Web Surveys”, *Public Opinion Quarterly* 72(5), 2008.

no fracasso das pesquisas de intenção. Para um artigo defendendo que a mentira pode ser um problema na tentativa de prever o apoio a Trump, veja Thomas B. Edsall, “How Many People Support Trump but Don’t Want to Admit It?” *New York Times*, 15 de maio de 2016, SR2. Mas para um defensor de que isto não foi um fator significativo, veja Andrew Gelman, “Explanations for That Shocking 2% Shift” Statistical Modeling, Causal Inference, and Social Science, 9 de novembro de 2016, http://andrewgelman.com/2016/11/09/explanations-shocking-2-shift/ [conteúdo em inglês].

explicou Tourangeau. Entrevistei Roger Tourangeau pelo telefone em 5 de maio de 2015.

tantas pessoas dizem estar acima da média. Isto é discutido no artigo de Adam Grant, *Originals: How Non-Conformists Move the World* (Nova York: Viking, 2016). A fonte primária é o artigo de David Dunning, Chip Heath e Jerry M. Suls, “Flawed Self-Assessment: Implications for Health, Education, e the Workplace” *Psychological Science in the Public Interest* 5 (2004).

atrapalhar a pesquisa. Anya Kamenetz, “ ‘Mischievous Responders’ Confound Research on Teens”, *nprED*, 22 de maio de 2014, http://www.npr.org/sections/ed/2014/05/22/313166161/mischievous-responders-confound-research-on-teens [conteúdo em inglês]. A pesquisa original que esse artigo discute é de Joseph P. Robinson-Cimpian, “Inaccurate Estimation of Disparities Due to Mischievous Responders” *Educational Researcher* 43, n° 4 (2014).

buscam mais pela palavra “pornô” do que por “clima”. https://www.google.com/trends/explore?date=all&geo=US&q=porn,weather [conteúdo em inglês].

admitem assistir à pornografia. Amanda Hess, “How Many Women Are Not Admitting to Pew That They Watch Porn?” *Slate*, 11 de outubro de 2013, http://www.slate.com/blogs/xx_factor/2013/10/11/pew_online_viewing_study_percentage_of_women_who_watch_online_porn_is_growing.html [conteúdo em inglês].

“pinto”, “foda” e “pornô”. Nicholas Diakopoulus, “Sex, Violence, e Autocomplete Algorithms” *Slate*, 2 de agosto de 2013, http://www.slate.com/articles/technology/future_tense/2013/08/words_banned_from_bing_and_google_s_autocomplete_algorithms.html [conteúdo em inglês].

são 3,6 vezes mais propensos a dizer ao Google que se arrependem. Estimo que, incluindo diversas construções de frases, há cerca de 1.730 buscas no Google por norte-americanos todo mês dizendo explicitamente que se arrependem de ter filhos. Há apenas 50 expressando arrependimento por não os ter. Existem cerca de 15,9 milhões de norte-americanos com mais de 45 anos que não têm filhos. Há cerca de 152 milhões que os têm. Isto significa, entre a população elegível, que pessoas com filhos têm 3,6 vezes mais probabilidade de expressar arrependimento no Google do que pessoas sem filhos. Obviamente, como mencionado no texto, mas digno de ênfase, estas confissões ao Google são feitas apenas por um pequeno e seleto número de pessoas — presumivelmente este é um sentimento forte o bastante para que as pessoas se esqueçam momentaneamente de que o Google não pode ajudá-las.

maior apoio ao casamento gay. Estas estimativas são do artigo de Nate Silver, “How Opinion on Same-Sex Marriage Is Changing, e What It Means” FiveThirtyEight, 26 de março de 2013, http://fivethirtyeight.blogs.nytimes.com/2013/03/26/how-opinion-on-same-sex-marriage-is-changing-and-what-it-means/?_r=0 [conteúdo em inglês].

Cerca de 2,5% dos usuários homens do Facebook que indicam o gênero em que estão interessados dizem que preferem homens. Análise do autor dos dados de anúncios do Facebook. Não incluo os dados de usuários do Facebook que indicam “homens e mulheres”. Minha análise sugere que um porcentual incomum de usuários que dizem estar interessados em homens e mulheres interpreta a pergunta como interesse em amizade e não romântico.

cerca de 5% das buscas de homens por pornografia são por pornografia masculina gay. Conforme discutido, o Google Trends não discrimina as buscas por gênero. O Google AdWords separa as visualizações de páginas em diversas categorias por gênero. Entretanto, estes dados são bem menos precisos. Para estimar as buscas por gênero, primeiro usei os dados de busca para obter uma estimativa dentro do estado do porcentual de buscas por pornografia gay por estado. Depois, padronizo os dados com os dados de gênero do Google AdWords. Outra maneira de obter dados específicos por gênero é usando os do PornHub. Entretanto, o PornHub poderia ser uma amostra altamente selecionada, já que muitos gays podem preferir usar sites especializados em pornografia gay. O PornHub sugere que o consumo de pornografia gay entre os homens é menor do que as buscas no Google sugerem. Entretanto, isto confirma que não há uma forte relação entre a tolerância em relação à homossexualidade e ao consumo de pornografia gay. Todos esses dados e observações adicionais estão disponíveis em meu site, em sethsd.com [conteúdo em inglês], na seção “Sex”.

4% deles são abertamente gays no Facebook. Cálculo do autor dos anúncios do Facebook. Em 8 de fevereiro de 2017, aproximadamente 300 estudantes do sexo masculino do ensino médio no mercado de mídia em San Francisco, Oakland San Jose, no Facebook disseram se interessar por homens. Aproximadamente 7.800 disseram ter interesse por mulheres.

“No Irã não temos homossexuais…”. “‘We Don’t Have Any Gays in Iran’, Iranian President Tells Ivy League Audience”, Daily Mail.com, 25 de setembro de 2007, http://www.dailymail.co.uk/news/article-483746/We-dont-gays-Iran-Iranian-president-tells-Ivy-League-audience.html [conteúdo em inglês].

“Não temos gays em nossa cidade”. Brett Logiurato, “Sochi Mayor Claims There Are No Gay People in the City”, *Sports Illustrated*, 27 de janeiro de 2014.

comportamento na internet revela significativo interesse por pornografia gay em Sochi e no Irã. De acordo com o Google AdWords, existem dezenas de milhares de buscas todo ano por “гей порно” (pornô gay). O porcentual das buscas pornográficas por pornô gay é praticamente igual em Sochi e nos Estados Unidos. O Google AdWords não inclui dados do Irã. O PornHub também não informa os dados para o Irã. Entretanto, o PornMD estudou estes dados de busca e relatou que cinco dentre os dez termos mais buscados no Irã eram por pornografia gay. As buscas incluíam “amor do papai” e “homem de negócios no hotel”, e são relatadas no artigo de Joseph Patrick McCormick, “Survey Reveals Searches for Gay Porn Are Top in Countries Banning Homosexuality”, *PinkNews*, http://www.pinknews.co.uk/2013/03/13/survey-reveals-searches-for-gay-porn-are-top-in-countries-banning-homosexuality/ [conteúdo em inglês]. De acordo com o Google Trends, cerca de 2% das buscas pornográficas no Irã são por pornô gay, o que é inferior aos Estados Unidos, mas ainda sugere um interesse disseminado.

Quando se trata de sexo. Stephens-Davidowitz, “Searching for Sex”. Dados para esta seção estão em meu site, sethsd.com [conteúdo em inglês], na seção “Sex”.

11% das mulheres. Atual Estado Contraceptivo entre as Mulheres de 15–44 anos: Estados Unidos, 2011–2013, Centers for Disease Control e Prevention, http://www.cdc.gov/nchs/data/databriefs/db173_table.pdf#1 [conteúdo em inglês].

10% delas ficassem grávidas todo mês. David Spiegelhalter, “Sex: What Are the Chances?” BBC News, 15 de março de 2012, http://www.bbc.com/future/story/20120313-sex-in-the-city-or-elsewhere [conteúdo em inglês].

1 para cada 113 mulheres em idade fértil. Ocorrem aproximadamente 6,6 milhões de gravidezes a cada ano e existem 62 milhões de mulheres entre os 15 e os 44 anos.

a performance no sexo oral no sexo oposto. Como mencionado, não sei o gênero de uma pessoa que faz uma busca no Google. Estou presumindo que a imensa maioria de buscas por como fazer sexo oral em mulheres é feita por homens e vice-versa. Isto porque a grande maioria das pessoas é heterossexual e porque não há necessidade de aprender a satisfazer um parceiro do mesmo sexo.

as cinco principais palavras negativas. A análise do autor nos dados do Google AdWords.

matar muçulmanos. Evan Soltas e Seth Stephens-Davidowitz, “The Rise of Hate Search”, *New York Times*, 13 de dezembro de 2015, SR1. Dados e mais detalhes podem ser encontrados em meu site, sethsd.com, na seção “Islamophobia”.

dezessete vezes mais comuns. A análise do autor dos dados do Google Trends.

Dia de Martin Luther King Jr. Análise do autor dos dados do Google Trends.

se correlaciona com a diferença de salário entre negros e brancos. Ashwin Rode e Anand J. Shukla, “Prejudicial Attitudes e Labor Market Outcomes”, mimeo, 2013.

Os pais delas. Seth Stephens-Davidowitz, “Google, Tell Me. Is My Son a Genius?” *New York Times*, 19 de janeiro de 2014, SR6. Os dados para as buscas exatas são encontrados usando o Google AdWords. Estimativas, com o Google Trends, comparando-se as buscas com as palavras “gênio” e “filho” versus “gênio” e “filha.” Compare, por exemplo, https://www.google.com/trends/explore?date=all&geo=US&q=gifted%20son,gifted%20daughter e https://www.google.com/trends/explore?date=all&geo=US&q=overweight%20son,overweight%20daughter [conteúdo das buscas pelos termos em inglês]. Uma exceção para o padrão geral de que há mais perguntas sobre a inteligência dos filhos e sobre o corpo das filhas são as buscas por “filho gordo”, que aparecem mais do que por “filha gorda”. Isto parece estar relacionado com a popularidade de pornografia incestuosa discutida anteriormente. Aproximadamente 20% das buscas com as palavras “gordo” e “filho” também incluem a palavra “pornô”.

meninas têm 9% mais probabilidade do que meninos de entrar em programas para crianças superdotadas. “Gender Equity in Education: A Data Snapshot”, Gabinete dos Direitos Civis, Departamento de Educação dos Estados Unidos, junho 2012, http://www2.ed.gov/about/offices/list/ocr/docs/gender-equity-in-education.pdf [conteúdo em inglês].

Cerca de 28% das meninas e 35% dos meninos estão acima do peso. Data Resource Center for Child and Adolescent Health, http://www.childhealthdata.org/browse/survey/results?q=2415&g=455&a=3879&r=1 [conteúdo em inglês].

perfis do Stormfront. Stephens-Davidowitz, “The Data of Hate”. Os dados pertinentes podem ser baixados em sethsd.com [conteúdo em inglês], na seção de dados intitulada “Stormfront”.

Stormfront durante a candidatura de Donald Trump. As buscas no Google pelo Stormfront foram semelhantes em outubro de 2016 aos níveis de outubro de 2015. Este é um enorme contraste com a situação durante a primeira eleição de Obama. Em outubro de 2008, o interesse nas buscas por Stormfront aumentou quase 60% comparado ao mês de outubro anterior. No dia seguinte à eleição de Obama, as buscas no Google por Stormfront aumentaram aproximadamente dez vezes. No dia seguinte à vitória de Trump, as buscas por Stormfront aumentaram cerca de 2,5 vezes. Isto equivale aproximadamente ao aumento no dia seguinte da eleição de George W. Bush em 2004, e reflete em grande parte o interesse por notícias entre os viciados em notícias.

segregação política na internet. Matthew Gentzkow e Jesse M. Shapiro, “Ideological Segregation Online and Offline”, *Quarterly Journal of Economics* 126, n° 4 (2011).

amigos no Facebook. Eytan Bakshy, Solomon Messing e Lada A. Adamic, “Exposure to Ideologically Diverse News and Opinion on Facebook”, *Science* 348, n° 6239 (2015). Eles descobriram que, entre os 9% dos usuários ativos do Facebook que declaram sua ideologia, cerca de 23% de seus amigos que também declaram tem uma ideologia oposta, e 28,5% das notícias que visualizam no Facebook são de uma ideologia diferente. Estes números não são diretamente comparáveis com outros sobre segregação, pois incluem apenas uma pequena amostra de usuários do Facebook que declaram uma ideologia. Presumivelmente, estes usuários têm propensão muito maior de ser politicamente engajados e se associar com outros usuários politicamente ativos com a mesma ideologia. Se isto estiver correto, a diversidade entre todos os usuários pode ser muito maior.

como o Facebook. Outro fator que torna a mídia social surpreendentemente diversificada é que oferece uma grande vantagem para artigos extremamente populares e amplamente compartilhados, não importa sua inclinação política. Veja Solomon Messing e Sean Westwood, “Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online”, 2014.

mais amigos no Facebook do que offline. Veja Ben Quinn, “Social Network Users Have Twice as Many Friends Online as in Real Life”, *Guardian*, 8 de maio de 2011. Este artigo discute um estudo de 2011 conduzido pelo Cystic Fibrosis Trust, que descobriu que o usuário médio de redes sociais tem 121 amigos online e 55 físicos. De acordo com um estudo de 2014 da Pew Research, o usuário médio do Facebook tinha mais de 300 amigos. Veja Aaron Smith, “6 New Facts About Facebook”, 3 de fevereiro de 2014, http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/ [conteúdo em inglês].

laços de amizade frágeis. Eytan Bakshy, Itamar Rosenn, Cameron Marlow e Lada Adamic, “The Role of Social Networks in Information Diffusion”, *Proceedings of the 21st International Conference on World Wide Web*, 2012 [conteúdo em inglês].

“As previsões sombrias não se tornam realidade”. “Study: Child Abuse on Decline in U.S.”, Associated Press, 12 de dezembro de 2011.

o abuso infantil realmente diminuiu. Veja Seth Stephens-Davidowitz, “How Googling Unmasks Child Abuse”, *New York Times*, 14 de julho de 2013, SR5, e Seth Stephens-Davidowitz, “Unreported Victims of an Economic Downturn”, mimeo, 2013.

enfrentar longas esperas e acabam desistindo.“Stopping Child Abuse: It Begins With You”, *The Arizona Republic*, 26 de março de 2016.

maneiras não oficiais de interromper uma gravidez. Seth Stephens-Davidowitz, “The Return of the D.I.Y. Abortion”, *New York Times*, 6 de março de 2016, SR2. Dados e mais detalhes estão em meu site, sethsd.com [conteúdo em inglês], na seção“Self-Induced Abortion”.

circulação média similar. Alliance for Audited Media, Consumer Magazines, http://abcas3.auditedmedia.com/ecirc/magtitlesearch.asp [conteúdo em inglês].

no Facebook. Cálculos do autor, em 4 de outubro de 2016, usando o Gerenciador de anúncios do Facebook.

dez sites mais visitados. “List of Most Popular Websites”, Wikipédia. De acordo com Alexa, que rastreia o comportamento de navegação, desde 4 de setembro de 2016, o site pornográfico mais popular era o XVideos, e ele era o 57° site mais popular. De acordo com o SimilarWeb, desde 4 de setembro de 2016, o site pornográfico mais popular era o XVideos, e ele era 17° mais popular. Os dez mais populares, segundo Alexa, são Google, YouTube, Facebook, Baidu, Yahoo!, Amazon, Wikipedia, Tencent QQ, Google India e Twitter.

Na manhã do dia 5 de setembro de 2006. Esta história é do livro de David Kirkpatrick, *O Efeito Facebook: Os bastidores da história da empresa que conecta do mundo* (edição original: Nova York, Simon & Schuster, 2010).

ótimos negócios são construídos sobre segredos. Peter Thiel e Blake Masters, *De Zero a Um: O que aprender sobre o empreendedorismo com o Vale do Silício* (edição original: Nova York, The Crown Publishing Group, 2014).

diz Xavier Amatriain. Entrevistei Xavier Amatriain por telefone em 5 de maio de 2015.

principais perguntas que norte-americanos fizeram durante o discurso sobre o Estado da União de 2014 de Obama. Análise do autor dos dados do Google Trends.

desta vez em uma mesquita. “The President Speaks at the Islamic Society of Baltimore”, vídeo do YouTube postado em 3 de fevereiro de 2016, https://www.youtube.com/watch?v=LRRVdVqAjdw [conteúdo em inglês].

buscas cheias de ódio e raiva contra os muçulmanos despencaram nas horas seguintes ao discurso do ex-presidente. Análise do autor dos dados do Google Trends. Buscas por “mate os muçulmanos” foram menores do que no período de comparação de uma semana antes. Além disto, as buscas incluindo “muçulmanos” e uma das cinco palavras negativas mais populares sobre este grupo foram menores.

CAPÍTULO 5: AJUSTANDO O FOCO

como as experiências da infância influenciam o time de beisebol para o qual você vai torcer. Seth Stephens-Davidowitz, “They Hook You When You’re Young”, *New York Times*, 20 de abril de 2014, SR5. Dados e códigos para este estudo podem ser encontrados em meu site, sethsd.com [conteúdo em inglês], na seção “Baseball”.

o ano mais importante. Yair Ghitza e Andrew Gelman, “The Great Society, Reagan’s Revolution, e Generations of Presidential Voting”, manuscrito não publicado.

explica Chetty. Entrevistei Raj Chetty por telefone em 30 de julho de 2015.

escaparem do anjo da morte? Raj Chetty et al., “The Association Between Income e Life Expectancy in the United States, 2001–2014”, *JAMA* 315, n° 16 (2016).

Comportamento contagiante pode impulsionar esse efeito. Julia Belluz, “Income Inequality Is Chipping Away at Americans’ Life Expectancy”, vox.com, 11 de abril de 2016.

por que algumas pessoas mentem em seus impostos. Raj Chetty, John Friedman e Emmanuel Saez, “Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings”, *American Economic Review* 103, n° 7 (2013).

Decidi baixar os dados da Wikipédia. Isto é do artigo de Seth Stephens-Davidowitz, “The Geography of Fame” *New York Times*, 23 de março de 2014, SR6. Dados podem ser encontrados em meu site, sethsd.com [conteúdo em inglês], na seção “Wikipedia Birth Rate, by County”. Pela ajuda com o download e a codificação dos condados de nascimento de cada entrada no Wikipédia, agradeço a Noah Stephens-Davidowitz.

uma cidade universitária de tamanho considerável. Para mais evidências sobre o valor das cidades, veja o livro de Ed Glaeser, *O Triunfo da Cidade* (edição original Nova York: Penguin, 2011). (Glaeser foi meu orientador na graduação.)

muitos exemplos da vida imitando a arte. David Levinson, ed., *Encyclopedia of Crime e Punishment* (Thousand Oaks, CA: SAGE, 2002).

sujeitos expostos a um filme violento reportam mais raiva e hostilidade. Craig Anderson et al., “The Influence of Media Violence on Youth”, Psychological Science in the Public Interest 4 (2003).

os finais de semana em que o filme mais popular era violento. Gordon Dahl e Stefano DellaVigna, “Does Movie Violence Increase Violent Crime?” *Quarterly Journal of Economics* 124, n° 2 (2009).

As buscas no Google também são discriminadas minuto a minuto. Seth Stephens-Davidowitz, "Days of Our Digital Lives", New York Times, 5 de julho de 2015, SR4.

o álcool é um grande contribuinte para o crime. Anna Richardson e Tracey Budd, "Young Adults, Alcohol, Crime e Disorder" *Criminal Behaviour e Mental Health* 13, n° 1 (2003); Richard A. Scribner, David P. MacKinnon e James H. Dwyer, "The Risk of Assaultive Violence e Alcohol Availability in Los Angeles County", *American Journal of Public Health* 85, n° 3 (1995); Dennis M. Gorman, Paul W. Speer, Paul J. Gruenewald e Erich W. Labouvie, "Spatial Dynamics of Alcohol Availability, Neighborhood Structure e Violent Crime" *Journal of Studies on Alcohol* 62, n° 5 (2001); Tony H. Grubesic, William Alex Pridemore, Dominique A. Williams, e Loni Philip-Tabb, "Alcohol Outlet Density e Violence: The Role of Risky Retailers e Alcohol-Related Expenditures", *Alcohol e Alcoholism* 48, n° 5 (2013).

para justificar que seus quatro filhos joguem futebol americano. "Ed McCaffrey Knew Christian McCaffrey Would Be Good from the Start — 'The Herd,' "vídeo no YouTube, postado em 3 de dezembro de 2015, https://www.youtube.com/watch?v=boHMmp7DpX0 [conteúdo em inglês].

a analisar pilhas de dados. Pesquisadores descobriram mais utilizando dados de crimes desmembrados em pequenos períodos de tempo. Um exemplo? Denúncias de violência doméstica aumentam imediatamente depois que o time de futebol americano da cidade perde um jogo em que é favorito. Veja David Card e Gordon B. Dahl, "Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior", *Quarterly Journal of Economics* 126, n° 1 (2011).

Simmons pode estar certo. Bill Simmons, "It's Hard to Say Goodbye to David Ortiz" ESPN.com, 2 de junho de 2009, http://www.espn.com/espnmag/story?id=4223584 [conteúdo em inglês].

como podemos prever o futuro desempenho de um jogador de beisebol? Isto é discutido no livro de Nate Silver, *O Sinal e o Ruído: Por que tantas previsões falham e outras não* (Intrínseca) (edição original: Nova York: Penguin, 2012).

"batedores fortões", de fato, em média, têm o ápice de desempenho mais jovens. Ryan Campbell, "How Will Prince Fielder Age?" 28 de outubro de 2011, http://www.fangraphs.com/blogs/how-will-prince-fielder-age/ [conteúdo em inglês].

os dúplices de Ortiz. Estes dados foram gentilmente cedidos por Rob McQuown, da Baseball Prospectus.

Kohane pergunta. Entrevistei Isaac Kohane por telefone em 15 de junho de 2015.

James Heywood é um empreendedor. Entrevistei James Heywood por telefone em 17 de agosto de 2015.

CAPÍTULO 6: O MUNDO TODO É UM LABORATÓRIO

27 de fevereiro de 2000. Esta história é discutida, entre outros lugares, no artigo de Brian Christian, "The A/B Test: Inside the Technology That's Changing the Rules of Business", *Wired*, 25 de abril de 2012, http://www.wired.com/2012/04/ff_abtesting/ [conteúdo em inglês].

Quando os professores recebiam, a ausência caía pela metade. Esther Duflo, Rema Hanna e Stephen P. Ryan, "Incentives Work: Getting Teachers to Come to School", *American Economic Review* 102, n° 4 (2012).

quando Bill Gates soube do trabalho de Duflo. Ian Parker, "The Poverty Lab", *New Yorker*, 17 de maio de 2010.

engenheiros da Google executaram 7 mil deles. Christian, "The A/B Test".

41 tons levemente diferentes de azul. Douglas Bowman, "Goodbye, Google", stopdesign, 20 de março de 2009, http://stopdesign.com/archive/2009/03/20/goodbye-google.html [conteúdo em inglês].

O Facebook hoje conduz. Eytan Bakshy, "Big Experiments: Big Data's Friend for Making Decisions", 3 de abril de 2014, https://www.facebook.com/notes/facebook-data-science/big-experiments-big-datas-friend-for-making-decisions/10152160441298859/ [conteúdo em inglês]. Fontes de informação sobre estudos farmacêuticos estão em "How many clinical trials are started each year?" Quora post, https://www.quora.com/How-many-clinical-trials-are-started-each-year [conteúdo em inglês].

Optimizely. Entrevistei Dan Siroker por telefone em 29 de abril de 2015.

rendendo aproximadamente US$60 milhões em doações adicionais. Dan Siroker, "How Obama Raised $60 Million by Running a Simple Experiment", Optimizely blog, 29 de novembro de 2010, https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/ [conteúdo em inglês].

testes A/B de manchetes do Boston Globe. Os testes A/B no *Boston Globe* e os resultados foram fornecidos para o autor. Alguns detalhes dos testes do *Globe* estão em "The Boston Globe: Discovering e Optimizing a Value Proposition for Content", Marketing Sherpa Video Archive, https://www.marketingsherpa.com/video/boston-globe-optimization-summit2 [conteúdo em inglês]. Eles incluem uma conversa gravada entre Peter Doucette do *Globe* e Pamela Markey no MECLABS.

diz Benson. Entrevistei Clark Benson por telefone em 23 de julho de 2015.

acrescentou uma seta apontando para a direita dentro de um quadrado. "Enhancing Text Ads on the Google Display Network", Inside AdSense, 3 de dezembro de 2012, https://adsense.googleblog.com/2012/12/enhancing-text-ads-on-google-display.html [conteúdo em inglês].

usuários do Google as criticaram. Veja, por exemplo, "Large arrows appearing in google ads — please remove", DoubleClick Publisher Help Forum, https://productforums.google.com/forum/#!topic/dfp/p_TRMqWUF9s.

a aumento dos vícios comportamentais na sociedade contemporânea. Adam Alter, Irresistible: The Rise of Addictive Technology e the Business of Keeping Us Hooked (Nova York: Penguin, 2017).

principais vícios relatados ao google em 2016. A análise do autor dos dados do Google Trends.

disse Levitt em uma palestra. Isto é discutido em um vídeo atualmente exibido na página do *Freakonomics*, de Harry Walker Speakers Bureau, http://www.harrywalker.com/speakers/authors-of-freakonomics/ [conteúdo em inglês].

anúncios de cerveja e de refrigerantes exibidos durante o Super Bowl. Wesley R. Hartmann e Daniel Klapper, "Super Bowl Ads", manuscrito não publicado, 2014.

um garoto cheio de espinhas de cuecas. Para uma forte defesa de que de fato vivemos em uma simulação de computador, veja "Are We Living in a Computer Simulation?", de Nick Bostrom, Philosophical Quarterly 53, n° 211 (2003).

Dos 43 presidentes norte-americanos. Equipe do Los Angeles Times, “U.S. Presidential Assassinations e Attempts”, Los Angeles Times, 22 de janeiro de 2012, http://timelines.latimes.com/us-presidential-assassinations-and-attempts/ [conteúdo em inglês].

Vejamos John F. Kennedy e Ronald Reagan. Benjamin F. Jones e Benjamin A. Olken, “Do Assassins Really Change History?”, *New York Times*, 12 de abril de 2015, SR12.

Kadyrov morreu. Um vídeo perturbador do ataque pode ser visto em “Parade surprise (Chechnya 2004)”, vídeo do YouTube, postado em 31 de março de 2009, https://www.youtube.com/watch?v=fHWhs5QkfuY [conteúdo em inglês].

Hitler tinha mudado sua agenda. Esta história também é discutida no artigo de Jones e Olken, “Do Assassins Really Change History?”

quando o líder de uma nação é assassinado. Benjamin F. Jones e Benjamin A. Olken, “Hit or Miss? The Effect of Assassinations on Institutions e War”, *American Economic Journal: Macroeconomics* 1, n° 2 (2009).

ganhar na loteria não. Esta questão é demonstrada no artigo de John Tierney, “How to Win the Lottery (Happily)”, *New York Times*, 27 de maio de 2014, D5. O artigo de Tierney discute os seguintes estudos: Bénédicte Apouey e Andrew E. Clark, “Winning Big but Feeling No Better? The Effect of Lottery Prizes on Physical e Mental Health”, *Health Economics* 24, n° 5 (2015); Jonathan Gardner e Andrew J. Oswald, “Money e Mental Wellbeing: A Longitudinal Study of Medium-Sized Lottery Wins” *Journal of Health Economics* 26, n° 1 (2007); e Anna Hedenus, “At the End of the Rainbow: Post-Winning Life Among Swedish Lottery Winners” manuscrito não publicado, 2011. O artigo de Tierney também comenta o famoso estudo de 1978 — Philip Brickman, Dan Coates e Ronnie Janoff-Bulman, “Lottery Winners e Accident Victims: Is Happiness Relative?” *Journal of Personality e Social Psychology* 36, n° 8 (1978) —, que descobriu baseado em uma minúscula amostra que ganhar na loteria não torna alguém feliz.

seu vizinho ganhar na loteria. Veja Peter Kuhn, Peter Kooreman, Adriaan Soetevent e Arie Kapteyn, “The Effects of Lottery Prizes on Winners e Their Neighbors: Evidence from the Dutch Postcode Lottery”, *American Economic Review* 101, n° 5 (2011), e Sumit Agarwal, Vyacheslav Mikhed e Barry Scholnick, “Does Inequality Cause Financial Distress? Evidence from Lottery Winners e Neighboring Bankruptcies”, artigo em trabalho, 2016.

vizinhos de ganhadores da loteria. Agarwal, Mikhed e Scholnick, “Does Inequality Cause Financial Distress?”

médicos podem ser motivados por incentivos monetários. Jeffrey Clemens e Joshua D. Gottlieb, “Do Physicians’ Financial Incentives Affect Medical Treatment e Patient Health?” American Econ*omic Review* 104, n° 4 (2014). Observe que estes resultados não significam que médicos são maus. Na verdade, os resultados seriam problemáticos se os procedimentos adicionais requisitados pelos quais recebem mais realmente salvassem vidas. Se este fosse o caso, isto significaria que os médicos precisariam receber o suficiente para que pedissem exames necessários para salvar vidas. Os resultados de Clemens e Gottlieb sugerem, em vez disso, que os médicos requisitam tratamentos imprescindíveis para salvar vidas não importando se recebem ou não a mais para isto. Para procedimentos que não ajudam muito, os médicos precisam receber mais. Outra maneira de dizer isto: os médicos não prestam atenção aos incentivos financeiros para questões de vida ou morte; mas prestam muita atenção àqueles de questões irrelevantes.

US$150 milhões. Robert D. McFadden e Eben Shapiro, “Finally, a Face to Fit Stuyvesant: A High School of High Achievers Gets a High-Priced Home”, *New York Times*, 8 de setembro de 1992.

Oferece. A oferta de cursos está disponível no site da Stuy, http://stuy.enschool.org/index.jsp [conteúdo em inglês].

um quarto dos formados é aceito. Anna Bahr, “When the College Admissions Battle Starts at Age 3”, *New York Times*, 29 de julho de 2014, http://www.nytimes.com/2014/07/30/upshot/when-the-college-admissions-battle-starts-at-age-3.html [conteúdo em inglês].

A Stuyvesant formou. Sewell Chan, “The Obama Team’s New York Ties”, *New York Times*, 25 de novembro de 2008; Evan T. R. Rosenman, “Class of 1984: Lisa Randall”, *Harvard Crimson*, 3 de junho de 2009; “Gary Shteyngart on Stuyvesant High School: My New York”, vídeo do YouTube, postado em 4 de agosto de 2010, https://www.youtube.com/watch?v=NQ_phGkC-Tk [conteúdo em inglês]; Candace Amos, “30 Stars Who Attended NYC Public Schools”, New York *Daily News*, 29 de maio de 2015.

Entre seus paraninfos estão. Carl Campanile, “Kids Stuy High Over Bubba: He’ll Address Ground Zero School’s Graduation”, *New York Post*, 22 de março de 2002; United Nations Press Release, “Stuyvesant High School’s ‘Multicultural Tapestry’ Eloquent Response to Hatred, Says Secretary-General in Graduation Address” 23 de junho de 2004; “Conan O’Brien’s Speech at Stuyvesant’s Class of 2006 Graduation in Lincoln Center”, vídeo do YouTube, postado em 6 de maio de 2012, https://www.youtube.com/watch?v=zAMkUE9Oxnc [conteúdo em inglês].

A Stuy é, em uma palavra, esplêndida. Veja https://k12.niche.com/rankings/public-high-schools/best-overall/ [conteúdo em inglês].

Menos de 5%. Pamela Wheaton, “8th-Graders Get High School Admissions Results” *Insideschools*, 4 de março de 2016, http://insideschools.org/blog/item/1001064-8th-graders-get-high-school-admissions-results [conteúdo em inglês].

das condições severas das prisões. M. Keith Chen e Jesse M. Shapiro, “Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach”, American Law and Economics Review 9, n° 1 (2007).

Os efeitos do Stuyvesant High School? Atila Abdulkadiroĝlu, Joshua Angrist e Parag Pathak, “The Elite Illusion: Achievement Effects at Boston e New York Exam Schools”, *Econometrica* 82, n° 1 (2014). O mesmo resultado foi encontrado de modo independente por Will Dobbie e Roland G. Fryer Jr., “The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools”, American Economic Journal: Applied Economics 6, n° 3 (2014).

um graduado mediano de Harvard ganha. *Veja* http://www.payscale.com/college-salary-report/bachelors [conteúdo em inglês].

alunos parecidos aceitos em instituições de prestígio semelhante que escolhem frequentar universidades diferentes acabam chegando praticamente ao mesmo patamar. Stacy Berg Dale e Alan B. Krueger, “Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables e Unobservables” *Quarterly Journal of Economics* 117, n° 4 (2002).

Warren Buffett. Alice Schroeder, *A Bola de Neve: Warren Buffett e o negócio da vida* [Sextante] (edição original: Nova York: Bantam, 2008).

CAPÍTULO 7: BIG DATA É CASCATA? O QUE ELE NÃO É CAPAZ DE FAZER

prever com sucesso quais as tendências de movimentação. Johan Bollen, Huina Mao, e Xiaojun Zeng, "Twitter Mood Predicts the Stock Market", *Journal of Computational Science* 2, n° 1 (2011).

O fundo hedge baseado em tuítes foi encerrado. James Mackintosh, "Hedge Fund That Traded Based on Social Media Signals Didn't Work Out" Financial Times, 25 de maio de 2012.

não conseguiram reproduzir a correlação. Christopher F. Chabris et al., "Most Reported Genetic Associations with General Intelligence Are Probably False Positives", *Psychological Science* (2012).

Zoë Chance. Esta história em TEDx Talks, "How to Make a Behavior Addictive: Zoë Chance at TEDx Mill River", vídeo do YouTube, postado em 14 de maio de 2013, https://www.youtube.com/watch?v=AHfiKav9fcQ [conteúdo em inglês]. Alguns detalhes da história, como a cor do pedômetro, foram desenvolvidos nas entrevistas. Entrevistei Chance por telefone em 20 de abril de 2015, e por e-mail em 11 de julho de 2016 e 8 de setembro de 2016.

Números são sedutores. Esta seção é do artigo de Alex Peysakhovich e Seth Stephens-Davidowitz, "How Not to Drown in Numbers", *New York Times*, 3 de maio de 2015, SR6.

trapaceou abertamente na administração dos testes. Brian A. Jacob e Steven D. Levitt, "Rotten Apples: An Investigation of the Prevalence e Predictors of Teacher Cheating", *Quarterly Journal of Economics* 118, n° 3 (2003).

diz Thomas Kane. Entrevistei Thomas Kane por telefone em 22 de abril de 2015.

"Cada medida acrescenta algo de valor". Bill e Melinda Gates Foundation, "Ensuring Fair e Reliable Measures of Effective Teaching", http://k12education.gatesfoundation.org/wp-content/uploads/2015/05/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf [conteúdo em inglês].

CAPÍTULO 8: MAIS DADOS, MAIS PROBLEMAS? O QUE NÃO DEVEMOS FAZER

Recentemente, três economistas. Oded Netzer, Alain Lemaire e Michal Herzenstein, "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications", 2016.

cerca de 13% dos tomadores. Peter Renton, "Another Analysis of Default Rates at Lending Club e Prosper" 25 de outubro de 2012, http://www.lendacademy.com/lending-club-prosper-default-rates/ [conteúdo em inglês].

curtidas no Facebook frequentemente se correlacionam. Michal Kosinski, David Stillwell, e Thore Graepel, "Private Traits and Attributes Are Predictable from Digital Records of Human Behavior", *PNAS* 110, n° 15 (2013).

acabam à mercê dos críticos do Yelp. Michael Luca, "Reviews, Reputation, and Revenue: The Case of Yelp" manuscrito não publicado, 2011.

Buscas no Google ligadas a suicídio. Christine Ma-Kellams, Flora Or, Ji Hyun Baek, e Ichiro Kawachi, "Rethinking Suicide Surveillance: Google Search Data and Self-Reported Suicidality Differentially Estimate Completed Suicide Risk", *Clinical Psychological Science* 4, n° 3 (2016).

3,5 milhões de buscas no Google. Este utiliza uma metodologia discutida em meu site nas notas sobre aborto autoinduzido. Comparo as buscas no Google da categoria "suicídio" com as buscas por "como dar nó em uma gravata". Houve 6,6 milhões de buscas no Google para "como dar nó em uma gravata" em 2015. Houve 6,5 vezes mais buscas na categoria suicídio. 6,5X6,6/12 » 3,5.

12 homicídios de muçulmanos reportados como crimes de ódio. Bridge Initiative Team, "When Islamophobia Turns Violent: The 2016 U.S. Presidential Election", 2 de maio de 2016, disponível em http://bridge.georgetown.edu/when-islamophobia-turns-violent-the-2016-u-s-presidential-elections/ [conteúdo em inglês].

CONCLUSÃO

O que motivou a cruzada de Popper? Karl Popper, *Conjecturas e Refutações* (edição original: Londres, Routledge & Kegan Paul, 1963).

mapeou todos os casos de cólera na cidade. Simon Rogers, "John Snow's Data Journalism: The Cholera Map That Changed the World", *Guardian*, 15 de março de 2013.

Benjamin F. Jones. Entrevistei Benjamin Jones por telefone em 1° de junho de 2015. Este trabalho também é discutido no artigo de Aaron Chatterji e Benjamin Jones, "Harnessing Technology to Improve K–12 Education", *Hamilton Project Discussion Paper*, 2012.

as pessoas tendem a não terminar obras de economistas. Jordan Ellenberg, "The Summer's Most Unread Book Is…" *Wall Street Journal*, 3 de julho de 2014.